Journal of Problem Solving

Human-Machine Cooperation in Large-Scale Multimedia Retrieval: A Survey

Kimiaki Shirahama,[1] Marcin Grzegorzek,[1] and Bipin Indurkhya[2]

[1]University of Siegen, [2]AGH University of Science and Technology

Correspondence:
Correspondence concerning this article should be addressed to Kimiaki Shirahama, Pattern Recognition Group, University of Siegen, Hoelderlinstrasse 3, 57076 Siegen, Germany, or via email to kimiaki.shirahama@uni-siegen.de.

Keywords:
large-scale multimedia retrieval, human-machine cooperation, machine-based methods, human-based methods

Large-Scale Multimedia Retrieval (LSMR) is the task to fast analyze a large amount of multimedia data like images or videos and accurately find the ones relevant to a certain semantic meaning. Although LSMR has been investigated for more than two decades in the fields of multimedia processing and computer vision, a more interdisciplinary approach is necessary to develop an LSMR system that is really meaningful for humans. To this end, this paper aims to stimulate attention to the LSMR problem from diverse research fields. By explaining basic terminologies in LSMR, we first survey several representative methods in chronological order. This reveals that due to prioritizing the generality and scalability for large-scale data, recent methods interpret semantic meanings with a completely different mechanism from humans, though such humanlike mechanisms were used in classical heuristic-based methods. Based on this, we discuss *human-machine cooperation*, which incorporates knowledge about human interpretation into LSMR without sacrificing the generality and scalability. In particular, we present three approaches to human-machine cooperation (*cognitive*, *ontological*, and *adaptive*), which are attributed to cognitive science, ontology engineering, and metacognition, respectively. We hope that this paper will create a bridge to enable researchers in different fields to communicate about the LSMR problem and lead to a ground-breaking next generation of LSMR systems.

INTRODUCTION

With the emergence of the Internet, the way to deliver visual and audio content has been significantly changed. The delivery in the early days was called *broadcasting*, where a small number of television and radio stations disseminated their programs to the general public. In the 90's, the distribution began to shift to *narrowcasting* through cable TV, Pay Per View (PPV), and so on. This enables the audience to select programs of interest from a much larger number of programs than the ones offered by broadcasting. Nowadays, the delivery is called *thincasting* (Snoek, & Smeulders, 2012) because video and audio hosting sites like YouTube and Internet Archive store a much higher number of programs compared to broadcasting and narrowcasting. For example, 300 hours of videos are uploaded to YouTube every minute (YouTube, n.d.). Such rapidly growing multimedia data cannot be manually managed or indexed.

It is often said that a picture is worth a thousand words. For example, the foreground video frame in Figure 1 conveys to humans many semantic meanings, such as "person," "road," "car," "tree," "building," "sky," "street," "daytime," and so on. In addition, the time dimension adds further meanings, like object actions, camera movements, things or people coming in and out of the screen, and so on. The sequence of video frames in Figure 1 shows "a person is walking," "the camera follows him," and "the road is out-of-frame in the end." Compared to this, actual multimedia data are indexed only with a small number of meanings. It is reported that on average, videos on YouTube are tagged only with one to seven meanings (Syrett, 2009). Therefore, a lot of research effort has been put on the development of *Large-Scale Multimedia Retrieval* (LSMR) methods, which analyze a large amount of multimedia data in terms of various semantic meanings, and support users to efficiently find interesting and relevant contents.

We adopt two policies in order to make the following discussions simple and clear. First, we use *example* to indicate a single unit of multimedia data, such as image, video, and audio. When the discrimination among these data formats is not important, we use examples as their abstract name. Second, by drawing an analogy with Content-Based Image Retrieval (CBIR) in Datta, Josh, Li, and Wang (2008), we define LSMR as any technology that, in principle, helps to organize large-scale multimedia data. Hence, LSMR in this paper includes technologies such as object recognition, image/video/audio classification, browsing, summarization, and so on.

The goal of LSMR is to quickly analyze a large amount of examples and accurately identify the ones relevant to a given query. In other words, LSMR can be considered as a binary classification problem to discriminate between relevant and

http://dx.doi.org/10.7771/1932-6246.1173

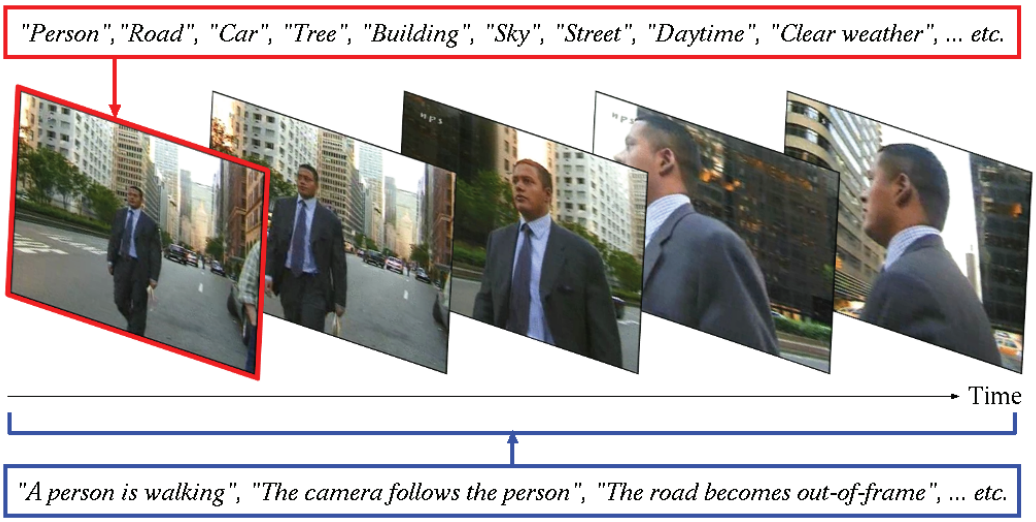

Figure 1. An illustration of various semantic meanings contained in multimedia data.

irrelevant examples to the query. It should be noted that there is a crucial difference between traditional alpha-numeric data and multimedia data (Shirahama, Ideno, & Uehara, 2006). The former are "structured" where their alpha-numeric representations directly describe semantic meanings and relationship operators (e.g., equal, not equal) are well-defined. On the other hand, raw multimedia data are "unstructured" where their digitized representations (i.e., pixel values on each image or video frame, and values in an audio signal) do not describe semantic meanings, and relationship operators are ill defined.

Thus, LSMR is generally conducted based on the scheme shown in Figure 2. From raw multimedia data, *features* that characterize meanings like color, edge, motion, and power spectrum are extracted at first. This is considered as the transformation of raw multimedia data into data that are computationally tractable and suitable for retrieval. Features are usually extracted as vectors because of their computational simplicity, and many sophisticated methods have been developed for vector data. For example, in Figure 2, the color feature representing many grey-colored pixels characterizes colors of the car and the road shown in the left video. In addition, the shape of the car and the boundary line of the road are characterized by the edge feature with many lines from top-left to bottom-right, and the movement of the car

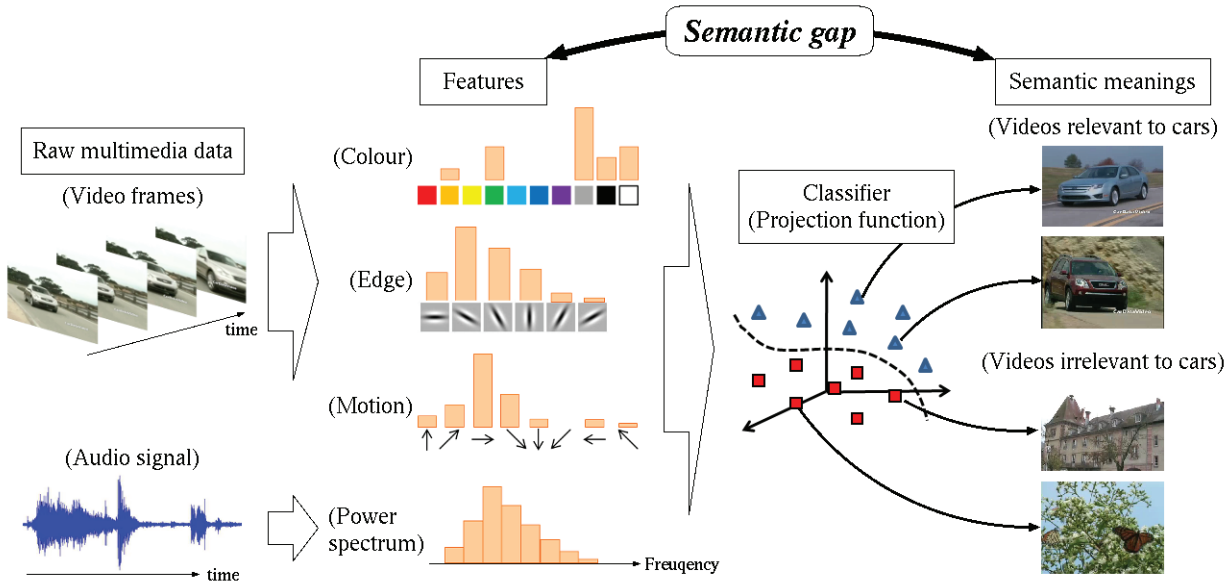

Figure 2. A general scheme of Large-Scale Multimedia Retrieval (LSMR).

is captured by the motion feature representing many right movements of (tracked) points. Also, the power spectrum indicates the main frequency of the engine sound of the car. By combining these features into a single vector, each example is represented as a point in a multidimensional space, as shown in the center of Figure 2. Based on this, a *classifier* (retrieval model) is constructed to discriminate between relevant and irrelevant examples to a query. In Figure 2, examples over and under the dashed line (classification boundary) are regarded as relevant and irrelevant to cars, respectively. Like this, the classifier can be considered as a projection function $f : \Re^n \rightarrow b$, where $\Re^n$ is the vector representation of an example and b is a binary variable representing the relevance or irrelevance to the query.

However, there is the *semantic gap*, which is the disagreement between features automatically extracted by machines and semantic meanings perceived by humans (Djordjevic, Izquierdo, & Grzegorzek, 2007; Smeulders, Worring, Santini, Gupta, & Jain, 2000; Staab et al., 2008). The semantic gap is attributed to the *internal dissimilarity* and the *external similarity* in terms of features. The former means that features in examples containing a certain meaning can significantly vary depending on camera techniques and shooting environments. The latter means that different meanings are often presented in examples with similar features. Let us consider examples displaying cars in Figure 3. As shown in Figure 3 (a), visual appearances (i.e., features) depend on different variables, such as the distance of the camera to a car, the shape of a car, lighting condition, and occlusion (other objects mask the shape of a car). In addition, Figure 3 (b) shows that visual appearances in examples displaying ships, trains, and helicopters are similar to those in examples displaying cars. Thus, research on LSMR mainly targets how to bridge the semantic gap by accurately covering internally dissimilar examples and excluding externally similar examples.

In the last two decades, many LSMR methods have been proposed and the retrieval performance has improved significantly. However, except for very specific problems like face detection, there is no practical LSMR method that can achieve accurate retrieval for various semantic meanings. One main reason is that, due to the large data size that is unmanageable by humans, researchers tend to leave LSMR just to machines. In other words, the enhancement of machine performance and the popularization of machine learning, data mining, and big data analysis caused the false expectation that machines fed with a large amount of examples can learn a classifier the way humans do. As a result, many recent methods do not consider any mechanism of how humans interpret semantic meanings. This paper emphasizes the importance of *human-machine cooperation* that incorporates the mechanism of human interpretation into LSMR. This approach complements the advantage of humans with the advantage of machines to create a synergy. On one hand, a human can easily recognize meanings in examples, but this is still difficult for a machine. On the other hand, the machine can analyze a large amount of examples much faster than the human. Therefore, by conceptualizing LSMR based on human-machine cooperation, we aim to achieve fast retrieval that can recognize meanings with the accuracy similar to human interpretation.

We review existing LSMR methods by classifying them into the following three categories: (1) *machine-based,* (2) *human-based,* and (3) human-machine cooperation. Machine-based LSMR does not explicitly model the mechanism of human interpretation. The most intuitive method is to construct a classifier only by statistically analyzing features of examples. Human-based LSMR is supported by humans, but machine and human are independent of each other. For example, human-based LSMR includes retrieval on multimedia data that are annotated by humans in advance, as current image and video hosting sites rely on manually provided text descriptions. Human-based LSMR also includes interactive approaches in which a classifier is iteratively refined by judging the relevance or irrelevance of currently retrieved examples. It should be noted that this interaction does not affect the algorithm of the classifier, but provides external data (i.e., judgement) to tune its parameters. Compared to this, human-machine cooperation addresses the collaboration of humans and machines at the algorithm level.

The survey approach of this paper is significantly different from those of existing papers in the field of multimedia processing. Most survey papers adopt a *progressive* approach to derive future research directions from the progress of component technologies. Specifically, recent survey papers

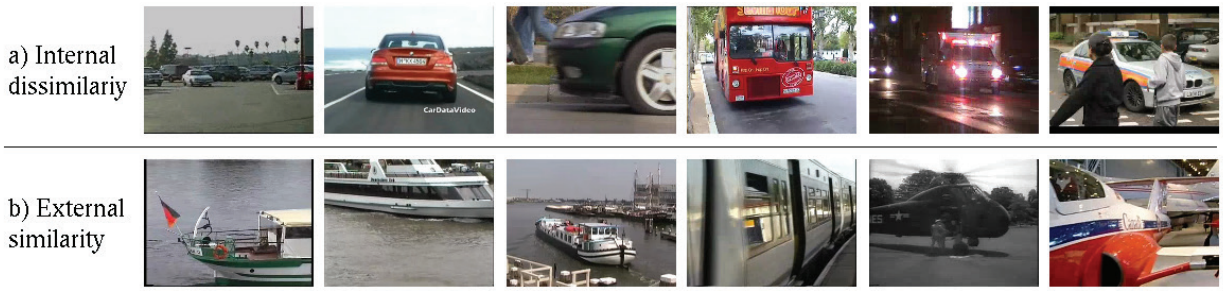

Figure 3. An example of the internal dissimilarity and the external similarity for examples displaying cars.

(e.g., Liu, Zhang, Lu, and Ma (2007), Bhatt and Kankanhalli (2011), Datta et al. (2008), Lew, Sebe, Djeraba, and Jain (2006), Snoek and Worring (2009), and Jiang, Bhattacharya, Chang, and Shah (2013)) mainly reviewed the following four component technologies: (1) feature extraction, representation, and transformation methods; (2) classifiers based on knowledge bases, machine learning techniques, similarities in terms of features, and data mining techniques; (3) user interaction methods such as query specification, browsing (visualization), and feedback; and (4) benchmark datasets for objectively evaluating the retrieval performance. Then, these papers suggest future problems that should be further explored or should receive more attention, such as improvement of component technologies, design of application-oriented (human-centric) interfaces, scalability with both high-performance computing and algorithm sophistication, synergy between different media like text, image, video, and audio, and utilization of user-generated web data like tagged images and videos.

Compared to such existing surveys, this paper conducts a survey taking a *retrospective* approach. By tracing the progress of LSMR from classical heuristic-based (or manual-based) approaches to recent machine learning-based (or web-based) approaches, we detect missing links from the latter, which were addressed by the former. That is, recent approaches consider knowledge about human interpretation of semantic meanings only to a certain degree, while it was fully used in classical approaches. Then, we discuss three directions of human-machine cooperation. The first one, based on knowledge about the human visual system, implements the mechanism of how human brains process visual information. The second direction, based on knowledge about human inference, effectively uses detectable semantic meanings (e.g., objects) to infer higher-level ones (e.g., events caused by objects' interaction). The last direction, based on knowledge about human learning, adaptively controls components of LSMR methods in an interactive process. To sum up, this paper advocates a "return" to the classical approaches, but we also need to consider much larger and much more structured knowledge in the future of LSMR.

Finally, this paper complements another survey paper that we have recently published, taking the above-mentioned retrospective approach in the field of multimedia processing (Shirahama & Grzegorzek, 2014). In that survey, we realized that the development of human-machine cooperation requires interdisciplinary expertise such as cognitive science, neuroscience, and ontology engineering. Consequently, this paper aims to disseminate the problem of human-machine cooperation in LSMR to many researchers in different fields to stimulate interdisciplinary collaborations. To this end, rather than covering various existing methods like Shirahama and Grzegorzek (2014), this paper concentrates on providing intuitive explanations of representative methods. Specifically, the next section focuses on three types of popular machine-based LSMR methods: classical methods using heuristically defined templates, methods that build classifiers using user-provided examples, and their extension in terms of features and classifiers. The following section addresses three types of standard human-based LSMR methods: classical methods based on manual annotation, their extension to the web-scale, and the most popular interactive methods based on user feedback. With respect to the material covered in these two sections, Shirahama and Grzegorzek (2014) assume familiarity with multimedia processing, presenting various machine-based and human-based methods using only one figure. In contrast, this paper graphically elaborates core ideas of all these methods (except intuitive manual annotation methods). In the LSMR based on Human-Machine Cooperation section, we also graphically illustrate methods that include not only notable utilization of knowledge about human interpretation, but also fundamental elements to understand the other methods described in Shirahama and Grzegorzek (2014). In addition, considering the research in cognitive science, this section provides several new ideas to achieve novel human-machine cooperation. Finally, readers can refer to Shirahama and Grzegorzek (2014) for diverse variants and extensions of methods described in this paper, and a detailed categorization and history of LSMR methods.

MACHINE-BASED LSMR

This section surveys machine-based LSMR methods. We firstly review classical heuristic approaches, and then present current popular machine learning approaches. Finally, we discuss the insufficiency of machine-based LSMR to bridge the semantic gap.

Note that most of the discussion in this section focuses on video data, but it is rather straightforward to apply it to image data. Before this discussion, let us define *shot* and *scene*, which are basic terminologies in video processing (Monaco, 1981). A shot is a sequence of video frames recorded continuously by a single camera. This is a basic physical unit where the content is spatially and temporally continuous. A scene is defined as a sequence of shots which are coherent to a certain semantic meaning such as location, action, or theme. For example, a conversation scene between two persons is presented by connecting a shot where one of the persons appears with a shot where the other person appears (see Figure 4). That is, scenes show higher-level semantic meanings than shots. In accordance with the definitions of shot and scene, we discuss machine-based LSMR methods below.

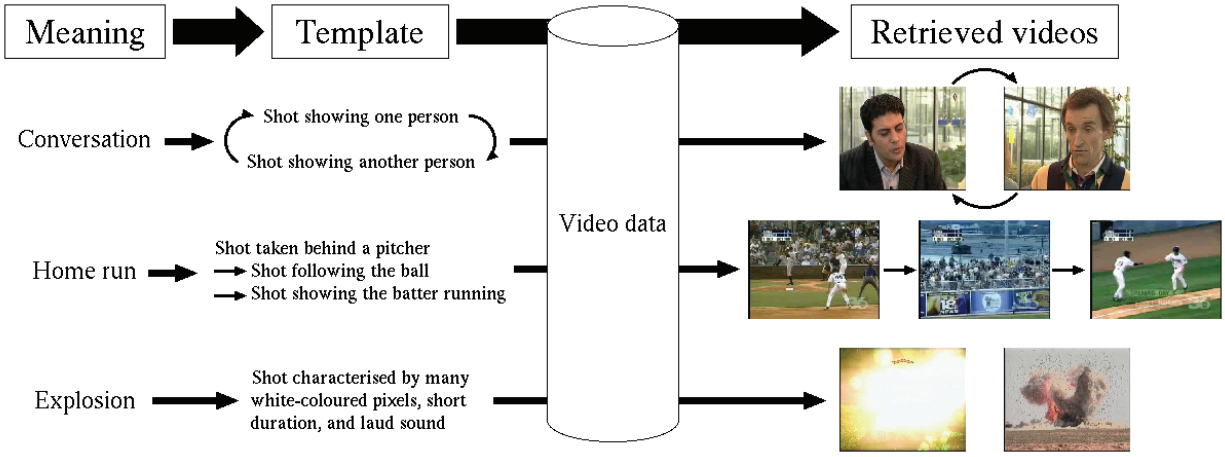

Figure 4. A general overview of heuristic approaches.

HEURISTIC APPROACHES

Classical heuristic approaches utilize prior knowledge about contents and structures of specific videos. A general overview of heuristic approaches is illustrated in Figure 4, where the key idea is to prepare *templates* that individually characterize a certain semantic meaning. For example, a conversation scene is characterized by a sequence of shots, where shots showing one person alternate with shots showing another person. A home run scene in a baseball video is presented by a shot sequence, where the first shot is taken behind the pitcher, the second shot follows the ball, and the third shot shows the batter running. Furthermore, a shot showing an explosion is marked by many white-colored pixels, a short duration, and a loud sound, because the explosion involves the flash, smoke, and explosive sound, and occurs in a moment. By preparing templates based on such prior knowledge, heuristic approaches retrieve shots or scenes which match those templates. Thus, templates work as classifiers in Figure 2.

The main research topic of heuristic approaches is the preparation of templates. We explain this below by using three examples in Figure 4. First, persons do not move in most conversation scenes. Thus, shots where each person appears have similar features (i.e., these shots are visually similar), so a template for conversation scenes is defined as the alternation between two types of shots where each type is characterized by similar features (Yoshitaka, Ishii, Hirakawa, & Ichikawa, 1997; Zhai, Rasheed, & Shah, 2004). To improve the retrieval performance, face detection developed in Viola and Jones (2001) is used in Zhai et al. (2004). Second, a baseball video is taken by a small number of cameras located at certain places in the stadium. This means that shots captured by one camera have similar features, and can be easily distinguished from shots captured by other cameras. Based on this, Chang, Han, and Gong (2002) and Ando, Shonida, Furui, and Mochizuki (2006) modeled a template for home run scenes as a Hidden Markov Model (HMM), where each state represents the probability distribution of features of shots taken by a camera, and such states are connected with transition probabilities. This approach can be used to prepare templates for other scenes like hit, catch, and infield play scenes. Last, Shirahama, Otaka, and Uehara (2007) represented a template for explosions as a logical conjunction of characteristic features regarding color, shot duration, and sound volume.

The aforementioned heuristic approaches can only process a limited number of a priori known queries. In contrast, users issue a variety of queries that cannot be assumed in advance. To overcome this, some research effort has been made on *video data mining*, where videos are analyzed using data mining techniques that extract previously unknown, interesting patterns in underlying data (Shirahama et al., 2006). This enables us to extract patterns for retrieving shots and scenes showing a variety of semantic meanings. We have developed a method that extracts sequential patterns for associating adjacent shots related to a certain meaning (Shirahama et al., 2006). Such sequential patterns are extracted by connecting statistically correlated features in temporally close shots. However, the extraction of sequential patterns is computationally expensive because numerous sequences of features have to be examined as candidates for patterns. Hence, time constraints, called "semantic event boundary" and "temporal locality," are adopted to eliminate many semantically irrelevant sequences of features. Our video data mining method extracted 16 patterns characterizing battle, hunting, explosion, indoor, outdoor, and so on (Shirahama et al., 2007).

However, heuristic approaches intrinsically have two critical problems. First, even using video data mining, it is practically impossible to prepare all patterns (templates) that can respond to a variety of queries issued by users. Second, templates which are defined by targeting specific videos lack

the generality. For example, all conversation scenes are not necessarily presented by the alternation between two types of visually similar shots in Figure 4. In addition, the template for home run scenes in Figure 4 targets baseball videos created by professional editors, but videos created by amateurs express these scenes in different forms. Moreover, other meanings may be displayed in shots that contain features represented by the template for explosions (e.g., these features are contained in shots showing snow and involving background music). Like this, predefined templates are not so useful for large-scale video data including various genres of videos. Thus, the research focus was shifted to machine learning approaches as described in the next section. For the first problem, these approaches construct a classifier on the fly, every time a query is issued by a user. The second problem can be alleviated by devising sophisticated features that have high discrimination powers as well as robustness to changes in visual appearances.

MACHINE LEARNING APPROACHES

Machine learning is a technique to construct a classifier using training examples, which are already labeled with classes, and predict classes of unknown test examples. This is applied to LSMR as *Query By Example* (QBE) (also called content-based retrieval) (Izquierdo, Chandramouli, Grzegorzek, & Piatrik, 2007; Petkovic & Jonker, 2002), where a user provides some examples to represent a query. That is, these are training examples labeled as relevant to the query. Then, a classifier is built to examine whether examples in the database (i.e., test examples) are relevant or irrelevant to the query. It should be noted that we consider QBE as a general approach that can be used for any example. Here, features are extracted directly from the example by applying physical metrics or mathematical transformations to pixels or audio signals. In other words, we do not consider approaches that use features obtained from external resources like closed captions, transcripts, and web documents, because they are available only for limited examples. (See Yan and Hauptmann (2007) for approaches based on external text resources.) Furthermore, QBE considered in this section is triggered only by a query represented with examples, and does not use any keyword query. In the Ontological Approaches section, we will present an approach that can be regarded as an extension of QBE and accepts a multimodal query specification by a combination of examples and keywords.

Figure 5 illustrates an overview of QBE where three examples are given for the query, "a person appears with a computer." Note that this textual description of the query is just a label for the sake of explanation, and only user-provided examples are used in QBE. From each example, features are extracted and organized into a vector. That is, the example is located in the multidimensional space as depicted by the arrows connected to the blue points in Figure 5. Similarly, test examples are represented as points in this multidimensional space, like white points in Figure 5. Under this setting, a classifier is constructed to distinguish test examples relevant to the query from the others. Intuitively, QBE retrieves test examples that are similar to training ones in terms of features, by assuming that examples with similar features display the same or similar semantic meanings.

Classical QBE methods use a nearest neighbor classifier that considers the similarity between training and test examples. More concretely, in Figure 5, test examples within the dashed circles are regarded as relevant to the query and retrieved. These kinds of classical QBE methods have been studied by assiduously addressing the following two research

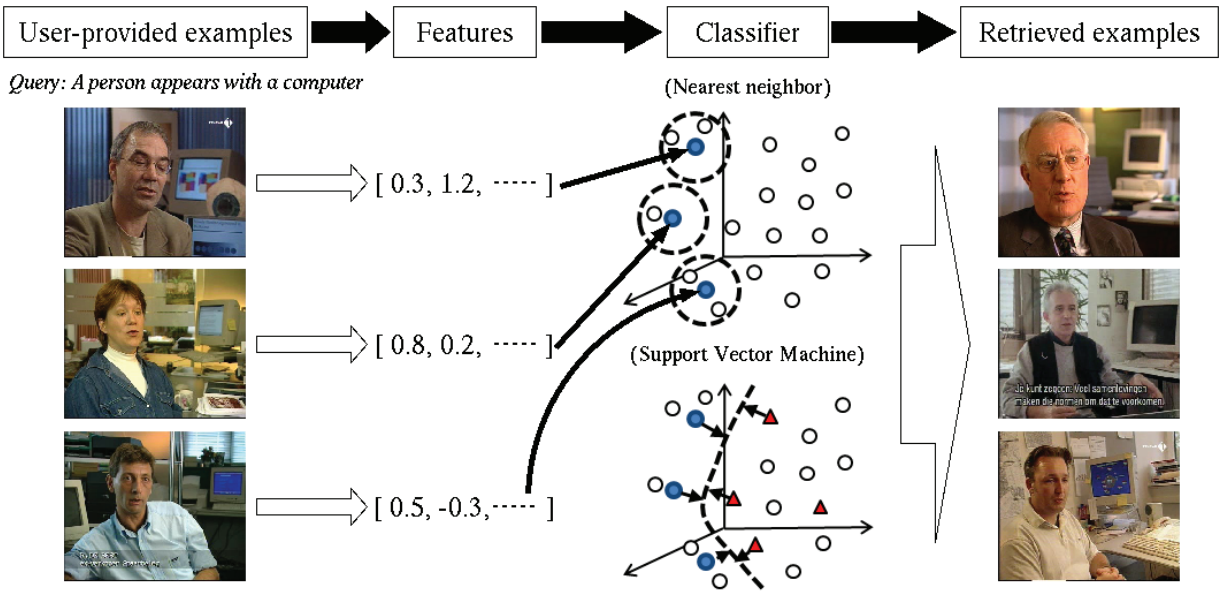

Figure 5. A general overview of Query By Example (QBE).

topics. The first is the development of good similarity measures between training and test examples. Many similarity measures such as a histogram-based measure (Jain, Vailaya, & Wei, 1999), psychology-based measure (Liu, Zhuang, & Pan, 1999), a measure based on weighted graph matching (Peng & Ngo, 2005), and a measure based on longest common subsequence (LCS) (Kim & Chua, 2005) have been developed. The other topic is the speed-up of the similarity calculation. For example, Kashino, Kurozumi, and Murase (2003) developed a method that avoids unnecessary similarity calculation by estimating the upper bound of similarity, and Yuan, Tian, and Ranganath (2004) devised a two-phase hierarchical method that first computes a coarse similarity on subsampled video frames, and then verifies the similarity using fine audio features.

However, classical QBE methods cannot achieve a satisfying retrieval accuracy. One reason is the weakness of *global features*, which are extracted from the whole region of an example. In other words, they only express overall characteristics of an example. As an example of global features, Figure 6 shows a color feature indicating the distribution of colors included in the example. This kind of overall representation loses a lot of information in an example. For instance, from the color feature in Figure 6, appearances of the car, road, and vegetation cannot be deduced any more. In addition, the overall characteristic of an example can easily change depending on camera techniques and shooting environments. For instance, the color feature of the example in Figure 6 changes substantially if it is taken in a brighter or darker lighting condition.

To overcome the weakness of global features, Schmid and Mohr (1997) proposed to represent an example as a collection of *local features*, each of which is extracted from a local region of the example. The top right of Figure 6 illustrates local features extracted from local regions, circled in yellow. In addition, Lowe (1999) developed a local feature called Scale-Invariant Feature Transform (SIFT), which represents the shape in a local region, reasonably invariant with respect to changes in illumination, rotation, scaling, and viewpoint. By extracting a large number of such local features from an example, we can ensure that at least some of them represent characteristic regions of a meaning. More specifically, even if the car in Figure 6 is partially masked by other objects, local features that characterize a wheel, window, or headlight are extracted from the visible part of the car.

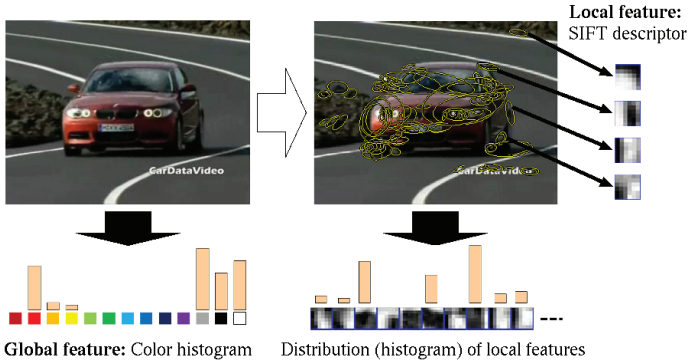

Figure 6. A comparison between a global feature and a local feature.

Based on local features, Csurka, Bray, Dance, Fan, and Williamowski (2004) developed a simple and effective example representation called *Bag of Visual Words* (BoVW), where each example is represented as the collection of characteristic local features, called *visual words*. In BoVW, millions of local features are first grouped into clusters where each cluster center is a visual word representing a characteristic local region. Then, each local feature extracted from an example is assigned to the most similar visual word. As a result, as seen from the bottom right of Figure 6, the example is represented as a vector (histogram) where each dimension represents the frequency of a visual word. This way, the example is summarized into a single vector where the detailed information is maintained by visual words (local features) that are robust with respect to varied visual appearances. The effectiveness of BoVW has been validated by many researchers (Csurka et al., 2004; Sande, Gevers, & Snoek, 2010; Jiang, Yang, Ngo, & Hauptmann, 2010; Shirahama, Matsuoka, & Uehara, 2012; Zhang, Marszalek, Lazebnik, & Schmid, 2007).

Another reason for the unsatisfactory performance of classical QBE methods is the insufficiency of training examples. A classification boundary between relevant and irrelevant examples to a query is supported only by training examples labeled as relevant (i.e., user-provided examples) (Juszczak & Duin, 2003). Below, for simplicity, we call these training examples *positive examples* because they serve as representatives of relevant examples to the query. Classical QBE methods just extract dense regions of positive examples in the multidimensional space. This requires a large number of positive examples to accurately shape regions of relevant examples to the query, but it is impractical for a user to provide many positive examples. Therefore, *negative examples*, which serve as representatives of irrelevant examples to the query, should also be used in QBE. Li and Snoek (2009) present that classifiers using both positive and negative examples are considerably superior to the ones only using positive examples.

However, a huge number of diverse examples can be negative because they only have to be irrelevant to a query. Thus, providing such negative examples is difficult for a user. With respect to this, Natsev, Naphade, and Tešić (2005) assumed that only a small number of examples in the database are relevant to the query, and all the others are irrelevant. Based on this, they proposed an approach that selects negative examples as randomly sampled examples because almost all of

them should be irrelevant to the query. This approach works reasonably well and has been utilized in many existing methods (Ngo et al., 2009; Snoek et al., 2009).

Using positive and negative examples, Natsev et al. (2005) proposed a QBE method that uses a Support Vector Machine (SVM) as a classifier. The SVM constructs a classification boundary based on the "margin maximization" principle so that it is placed in the middle between positive and negative examples. In other words, the distance (margin) of the boundary to the nearest positive (or negative) example is maximized (Vapnik, 1998). Figure 5 illustrates this margin maximization in the same multidimensional space to the nearest neighbor classifier. Here, in addition to three positive examples represented by blue points, four negative examples are selected as marked by red triangles. With the margin maximization, the classification boundary depicted by the dashed line is extracted by considering locations of three positive and three negative examples associated with arrows (one negative example is regarded as unnecessary). This "moderate" boundary, which is biased toward neither positive nor negative examples, is suitable for BoVW. Specifically, many visual words (i.e., thousands of visual words) are required to maintain the discrimination power of BoVW. That is, an example is represented as a high-dimensional vector. This renders the nearest neighbor classifier ineffective because of many irrelevant dimensions to similarity calculation. In contrast, the margin maximization makes the generalization error of an SVM independent of the number of dimensions, if this number is sufficiently large (Vapnik, 1998). Actually, SVMs have been successfully applied to BoVW with thousands of dimensions (Csurka et al., 2004; Jiang et al., 2010; Sande et al., 2010; Shirahama et al., 2012).

Note that any feature or classifier can be used in the framework of QBE in Figure 5. (See Jiang et al. (2013) for various global, local and audio features, and extensions of BoVW.) Regarding classifiers, although researchers have proposed many classifiers like tree-type classifiers, probabilistic classifiers and ensemble of classifiers (Bhatt & Kankanhalli, 2011; Jiang et al., 2013), SVM is currently considered as a standard classifier because of its simplicity and widely proven performance (Jiang et al., 2013; Snoek & Worring, 2009).

Below, we discuss machine learning approaches in object recognition on large-scale data. This can be formulated in the same way to QBE. That is, a classifier is built using positive and negative examples annotated with the presence or absence of a certain object, and is then used to distinguish test examples where this object appears from the rest of the test examples. However, object recognition needs to be performed on the "category level." (To be precise, this is called *generic object recognition*. In contrast, *specific object recognition* is the task of identifying the same instance of an object in different examples. Please refer to Grauman and Liebe (2011) for details of generic and specific object recognitions.) Although local features are useful for managing diverse visual appearances associated with the same or similar instances of the object, instances with significantly different appearances are included in the same object category. Taking Figure 3 (a) as an example, all the saloon cars, buses, and trucks should be recognized as cars. Regarding this, a classifier can conduct accurate recognition on test examples where instances of an object are similar to those in training examples. But recognition is not accurate on test examples where instances have significantly different characteristics from those in training examples. Thus, a large number of training examples are required to address the diversity attributed to the difference in instance types of an object. In general, the recognition performance is proportional to the logarithm of the number of positive examples, although each object has its own complexity of recognition (Naphade & Smith, 2004). This means that ten times more positive examples improve the performance by 10%. Considering the importance of the number of training examples, online systems for efficiently collecting large-scale training data have been developed where users on the web collaboratively annotate a large number of examples as positive or negative (Ayache & Quénot, 2008; Volkmer, Smith, & Natsev, 2005).

Another important issue for object recognition is the question of how to sample local features. In general, local feature extraction consists of two modules, *region detector* and *region descriptor* (Zhang et al., 2007). The former detects regions useful for characterizing objects, and the latter represents each of the detected regions as a vector. For example, SIFT features are typically extracted using a Harris-laplace (or Harris-affine) detector to identify regions where pixel values largely change in multiple directions. Such regions are regarded as useful for characterizing local shapes of objects, like corners of buildings, vehicles, and human eyes. Then, each detected region is described as a 128-dimensional vector representing the distribution of edge orientations. However, an object is shown in significantly different regions, and in videos, it does not necessarily appear in all video frames. Considering this "uncertainty" of object appearances, it is necessary to extract the BoVW representation of an example by exhaustively sampling local features in both the spatial and temporal dimensions. Actually, the performance is improved as the number of sampled local features increases (Nowak, Jurie, & Triggs, 2006). In addition, Snoek, Worring, Geusebroek, Koelma, and Senstra (2005) compared two methods. One extracts features only from one video frame in each shot (one shot contains more than 60 frames), and the other extracts features every 15 frames. They found out that the latter exceeds the former by 7.5 to 38.8%.

However, it requires expensive computational costs to process a large number of training examples and exhaustively sampled local features. So far, many methods for reducing

these computational costs have been developed based on special hardware like computer cluster (Yan, Fleury, Merler, Natsev, & Smith, 2009) and General-Purpose computing on Graphics Processing Units (GPGPU) (Sande, Gevers, & Snoek, 2011), or based on algorithm sophistication with sub-problem decomposition (Fan, Chen, & Lin, 2005) and tree structures (Inoue & Shinoda, 2012). In this context, we utilized matrix operations to develop a fast SVM training and test method, and a fast probabilistic BoVW extraction method (Shirahama & Uehara, 2012). The former reformulates similarity computation, which enables batch computation of similarities among many examples. The latter reformulates probability density computation, so that probability densities for many local features can be computed in a batch. Based on these, SVM training and test and BoVW extraction become about 10–37 and 5–7 times faster than the normal implementation, respectively. By processing a large number of training examples and exhaustively sampled local features using these methods, we achieved the highest performance in the TRECVID 2012 Semantic Indexing (light) task, which is one of the most famous worldwide competitions on object recognition (Shirahama & Uehara, 2012).

DISCUSSION

We have reviewed machine-based LSMR methods by putting them into heuristic and machine learning approaches. The above discussion reveals that, despite much research effort invested in machine learning approaches, the underlying framework remains the same. That is, a classifier is built by statistically analyzing locations of training examples in a multidimensional space defined by features. One main reason why this framework is favored is that researchers prioritize the generality and scalability, so that the same method can be used to search large-scale data for a variety of queries. However, we claim that this framework is limited because real-world examples are "unconstrained" in the sense that they can be taken by arbitrary camera techniques, and in arbitrary shooting environments (Jiang et al., 2013). Thus, a certain semantic meaning can be potentially associated with an infinite number of visual appearances that cannot be encompassed by training examples, even if they are provided in abundance as in object recognition. In other words, humans do not rely on a large number of training examples to interpret semantic meanings. Like this, the mechanism of recent LSMR methods prioritizing the generality and scalability has become completely different from the mechanism of human's semantic meaning interpretation.

By considering the chronological transition from heuristic to machine learning approaches, we can find that knowledge about human interpretation was utilized in the former, but was left out in the latter. With respect to this, the biggest disadvantage of heuristic approaches is that knowledge is represented just as a list of predefined templates. This clearly limits the generality to apply heuristic approaches to a variety of semantic meanings. In other words, each template is useful only for one meaning. Instead, we stress the necessity of knowledge that describes some "general" mechanism of how humans interpret semantic meanings. By utilizing such knowledge at the algorithm level in machine-based approaches, we expect that it is possible to achieve LSMR based on human interpretation without sacrificing the generality. We will discuss this in the LSMR based on Human-Machine Cooperation section.

HUMAN-BASED LSMR

This section first presents LSMR methods based on manual annotation of multimedia data. Then, we review approaches that enable users to interactively refine retrieval results. Finally, we discuss the problems of manual annotation and interactive approaches.

MANUAL ANNOTATION APPROACHES

These manual annotation approaches search over examples manually annotated with text descriptions. In the early years, this topic was investigated as a database problem to flexibly respond to various queries. In particular, video retrieval based on manual annotation was explored by addressing the following three issues (Tanaka, Ariki, & Uehara, 1999):

1. **Identification of meaningful segments:** Videos are known as *continuous media* where sequences of media quanta (i.e., video frames and audio samples) convey semantic meanings when continuously played over time (Gemmell, Vin, Kandlur, Rangan, & Rowe, 1995). Due to this temporal continuity, any segment of a video can become a meaningful unit.

2. **Annotation that should be provided:** A video contains many meanings ranging from primitive ones like color and shape to deep ones like story and event. It is difficult to annotate the video with all the semantic meanings contained in it.

3. **Discrepancy between annotation and user expectation:** This focuses on segments that are annotated and segments that are expected to be retrieved by users. Let us consider the query "person *A* and person *B* are talking to each other." One intuitive answer to this query is a shot that is annotated with both *A*'s and *B*'s presences. However, a sequence of shots can be another answer where shots annotated only with *A*'s presence and shots annotated only

with *B*'s presence are repeated one after the other. Thus, dynamic organization of annotated shots (segments) is required to correctly respond to queries.

In accordance with these issues, Oomoto and Tanaka (1993) developed Object-Oriented Video Information Database (OVID) where a segment and text descriptions are regarded as a video object and attribute values, respectively. Such attribute values of a video object are inherited by another object based on their temporal inclusion relationship. This way, text descriptions are shared among video objects so that manual annotation effort is significantly reduced. Uehara, Oe, and Maehara (1996) proposed an approach that represents the story of a video using a binary tree, called a story graph. In this graph, each node represents the relation (e.g., sequential, physically causal, and psychologically causal) between two successive segments, and edges are labeled with semantic constraints. This enables users to retrieve arbitrary-length scenes specified by natural language, and retrieve causes or consequences of queries based on causal relationships.

Pattanasri, Chatvichienchai, and Tanaka (2005) developed a video retrieval method using a knowledge base (ontology) about contexts. This knowledge base represents relationships among verbs, such as "kill" implies "die." Thereby, video segments that are related in terms of causes and effects of person's actions can be linked together and retrieved as a whole. François, Nevatia, Hobbs, Bolles, and Smith (2005) developed an extensible and hierarchical framework for representing events in videos. Here, complex events are constructed from simpler events by operations, such as sequencing, iteration, and alternation, which are defined in a knowledge base. Like this, various complex events can be defined only using relatively few primitive events.

Since manual annotation is a laborious task, the aforementioned approaches have the limitation in the scalability for large-scale data. Thus, they have been extended by distributing manual annotation of large-scale multimedia data to many users on the web. The following two issues are crucial for devising this web-based annotation:

1. **Usability:** This means whether users can easily annotate multimedia data or not. If this is insufficient, it cannot be expected that many users participate in annotation.

2. **Annotation quality:** When utilizing unfamiliar users on the web, meaningless annotation may be provided by malicious users or operation mistakes. In addition, different descriptions may be annotated to indicate the same meaning. For example, one user may annotate an example showing a car with the description "car," while it may be annotated with "automobile" by another user.

Considering these issues, Volkmer et al. (2005) developed a system for annotating a large number of shots with objects' presences or absences. To improve the usability, users are allowed to customize their annotation styles, such as the number, size, and layout of shots displayed per page, using mouse and/or keyboard, and annotating one or more objects at a time. In addition, the system informs the user how difficult the annotation of each object is based on the disagreement in past annotations by different users, so that the annotation quality is improved. Ayache and Quénot (2008) extended this system by combining manual annotation with shot selection based on active learning. Here, shots for which the recognition by the classifier of an object is the most uncertain are preferentially annotated by users. In other words, it is redundant to annotate shots for which the recognition seems confident, so that annotation cost can be significantly reduced. Russell, Torralba, Murphy, and Freeman (2008) developed LabelMe, a web-based system for annotating object regions in images. Given an image, a user labels an object region by creating a polygonal region by mouse, then types the object name. To improve the usability and maintain the annotation quality, the researchers considered several extensions, such as the lexical knowledge base (WordNet) for expanding and disambiguating freely typed object names, and the object relation for suggesting candidate objects where their regions frequently overlap a user-specified region.

Web-based annotation approaches described above have been further enhanced by considering the motivation of users. That is, regular users on the web are unlikely to volunteer to annotate when no benefit or no reason is given. To overcome this, Ahn and Dabbish (2004, 2008) proposed a *Games With a Purpose* (GWAP) approach where users play a game, and as a side effect, a computationally difficult task is solved. More concretely, users play a fun game without knowing that they conduct image annotation. Based on this idea, an ESP game is developed where randomly paired users are first given the same image, then each user guesses a label that another user is likely to provide (Ahn & Dabbish, 2004, 2008). If labels provided by both users agree, they get a certain number of points, and the next image is given. This way, users are encouraged to get more points and play the ESP game many times. Since users know nothing and cannot communicate with each other, the easiest way for them to earn points is to provide labels relevant to given images. Thus, annotations obtained by the ESP game are likely to be meaningful. The quality of annotation is further improved using taboo words that users are not allowed to type. As of July 2008, 200,000 users contributed to assigning more than 50 million labels to images on the web (Ahn & Dabbish, 2008). Several variants of the ESP game have been developed, such as games for object region annotations (Ahn, Liu, & Blum, 2006; Steggink & Snoek, 2011), video annotation

(Zwol, Garcia, Ramirez, Sigurbjornsson, & Labad, 2008), music annotation (Barrington, O'Malley, Turnbull, & Lanckriet, 2009) and geographically referenced photo annotation for landmark objects (Bell et al., 2009).

Another web-based annotation approach that motivates users is *crowdsourcing* which outsources problems performed by designated human (employee) to users on the web (Quinn & Bederson, 2011). In the field of multimedia processing, one of the most famous crowdsourcing systems is Amazon's Mechanical Turk, where anyone can post small tasks and specify prices paid for completing them (Kittur, Chi, & Suh, 2008). For example, Deng et al. (2009) used this to annotate 3.2 million images in terms of presences of 5,247 objects.

INTERACTIVE APPROACHES

We now focus on interactive approaches where a user iteratively refines the performance based on the current retrieval result. Interactive approaches are needed because of the *user individuality*, which means that even for the same query, different users may be interested in different examples (Zhou and Huang 2003). For example, for the query "horse," one user may look for examples showing "adult horse," while another may look for examples showing "child horse." In addition, it is often difficult for a user to precisely express his/her intent because of the poor lexical vocabulary or the lack of proper positive examples for QBE. For example, when the user wants to search for a specific model of a Porsche car, it often happens that he/she does not know the model name. Only specifying the keyword "Porsche" leads to retrieve examples showing different models. In the case of QBE, if a user queries a database for Barack Obama using a positive example showing him in front of a car, the retrieval result will contain not only examples where he appears, but also examples showing different cars. This is called the *intention gap*, which is the discrepancy between the user's search intent and the query specified by him/her (Zha et al,. 2010). Thus, the interactive refinement of retrieval results is necessary to overcome the user individuality and intention gap.

One of the most popular interactive approaches is *Relevance Feedback* (RF), which asks a user to provide feedback regarding the relevance or irrelevance of currently retrieved examples (Zhou & Huang, 2003). Using these newly annotated examples, the current classifier is refined. RF is closely related to *active learning* to select the most informative examples for improving the performance of a classifier, and asks the user to annotate them (Wang & Hua, 2011). Such RF (or active learning) methods enable us to achieve accurate retrieval with reduced manual annotation effort.

Figure 7 illustrates a typical RF based on an SVM. For the query "flowers," Figure 7 (a) shows a retrieval result where blue circles and red triangles are positive and

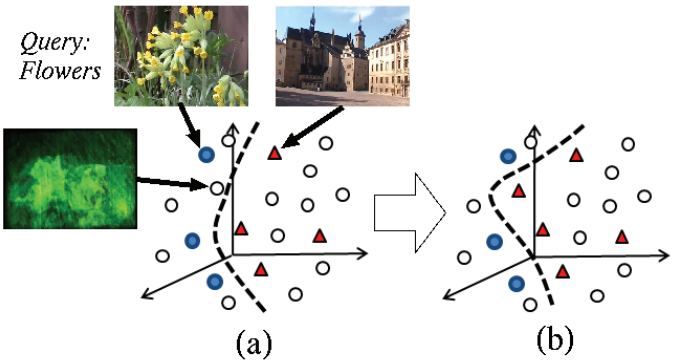

Figure 7. An illustration of Relevance Feedback (RF) based on an SVM.

negative examples to obtain the SVM's classification boundary depicted by the dashed line (see the Machine Learning Approaches section for the detail of SVM-based retrieval). Among test examples represented by circles, the ones in the left side of the boundary are currently retrieved. It is intuitive that the classification of the SVM is the most uncertain for the test example closest to the boundary (Tong & Chang, 2001; Wang & Hua, 2011). In Figure 7 (a), such a test example is indicated together with the image, which has the visual appearance like flowers, but does not show them. This test example is presented to the user and is annotated as negative. Using this as an additional training example, the boundary of the SVM is modified as shown in Figure 7 (b). Like this, the SVM is efficiently refined by asking the user to annotate the most uncertain test example. In other words, examples far from the boundary are regarded as being reasonably classified, thus labeling them is redundant. (See Wang and Hua (2011) and Shirahama and Grzegorzek (2014) for other types of RF approaches.)

DISCUSSION

We have reviewed two types of human-based LSMR: manual annotation and interactive approaches. The development of the former has both the strength and weakness. Compared to classical manual annotation approaches, recent web-based approaches are much more scalable for annotating large-scale data. However, the latter cannot offer flexible retrieval based on scenes and causal relations, while this was supported by the former based on annotation of deep semantic meanings. That is, recent web-based approaches have to make annotation simple in order to maintain the usability. In addition, GWAP approaches have a drawback in which users tend to maximize their scores, so collected descriptions only represent general properties of examples (e.g., color and shape) (Gupta, Li, Yin, & Han, 2010). Also, crowdsourcing requires huge monetary cost. To the best of our knowledge, web-based approaches only support annotation of primitive

semantic meanings like objects, and do not support annotation of deep ones like scenes and causal relationships. This chronological transition of manual annotation approaches is similar to the one of machine-based approaches, where knowledge about human interpretation used in classical approaches was left out in recent ones prioritizing the generality and scalability for large-scale data.

It should be noted that all the manual annotation approaches leave the most difficult tasks (i.e., interpretation of semantic meanings) to humans, and do not contribute to bridging the semantic gap. However, they have important roles in LSMR research. First, annotation obtained by web-based approaches such as those by Ayache and Quénot (2008), Deng et al. (2009), and Russakovsky et al. (2014) are recently used as training examples in machine learning approaches for object recognition. In addition, the final goal of human-machine cooperation, discussed in the next section, is to achieve automatic annotation of deep semantic meanings used in classical approaches.

Finally, interactive (RF) approaches somehow improve the retrieval performance using newly annotated examples as additional training examples. However, features and classifiers are substantially the same as those in machine learning approaches. In other words, interactive approaches just tune parameters of machine learning approaches. Instead, we argue that interactive approaches need to iteratively refine features and classifiers that are currently insufficient for representing complex semantic meanings. In the next section, we will discuss these approaches based on knowledge about human learning.

LSMR BASED ON HUMAN-MACHINE COOPERATION

In this section, we discuss human-machine cooperation methods by putting them into three categories: *cognitive*, *ontological*, and *adaptive*. We relate these categories as illustrated in Figure 8. First, cognitive methods utilize knowledge about the human visual system, where functionalities of the human brain are modeled to detect primitive semantic meanings like objects, and concepts defined in the Ontological Approaches section. The arrow (1) in Figure 8 represents that ontological methods using knowledge about human inference detect high-level semantic meanings based on relations of primitive ones detected by cognitive methods. On the other hand, the arrow (2) indicates that these relations can be used to validate and refine detection results by cognitive methods. The two arrows marked with (3) in Figure 8 present that adaptive methods based on knowledge about human learning take as input the information of features and classifiers in cognitive and ontological methods (metalevel features defined in the Adaptive Approaches section). The arrows denoted by (4) indicate that these features

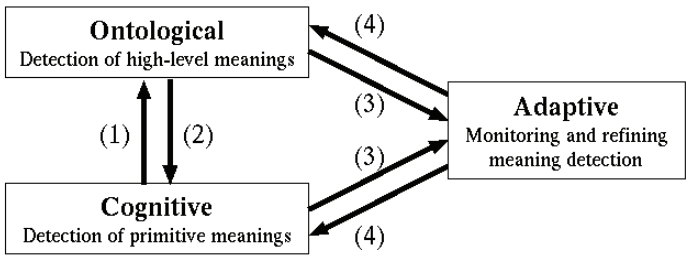

Figure 8. An illustration of the relation among cognitive, ontological, and adaptive approaches.

and classifiers are adaptively refined based on user feedback. Below, for each of the categories, we firstly describe the existing methods and then discuss how to extend them in the future.

COGNITIVE APPROACHES

Existing Approaches. *Cognitive science* is an interdisciplinary study of mind and intelligence in order to theoretically explain how the human mind (thinking) works (Ogiela & Tadeusiewicz, 2010; Pizlo, 2010, 2014). In other words, cognitive science tries to grasp the complex human mind by utilizing methods in different research fields, such as philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology. In particular, owing to psychological and neurological experiments, the human visual system is intensively investigated from the "sensation" process, which transduces the light (stimulus) received by the eye into neural signals, to the "perception" process, which translates neural signals into meanings. Thus, incorporating methods and knowledge in cognitive science into LSMR is beneficial for bridging the semantic gap.

When recognizing an object, humans are known to use two processes, *bottom-up* and *top-down*. The former process is driven by stimuli acquired from the external environment. More specifically, the bottom-up process starts with grouping visual attributes (e.g., color, brightness, and texture) in examples to form homogeneous regions. Typically, this does not provide an accurate result where the entire region of the object is fragmented into small regions due to various changing factors, such as camera technique, lighting condition, object shape deformation, and occlusion. On the other hand, the top-down process is driven by prior knowledge and expectations in the mind. An example of prior knowledge is the contour of the object. In addition, it has been empirically proven that humans use the symmetrical property of an object as prior knowledge, so that the 3D shape of this object can be efficiently recovered from its 2D appearance

in an image (Pizlo, 2014). However, the top-down process is difficult to conduct under the condition where the appearance of the object is vague due to various changing factors. Therefore, an *intermediate representation* is necessary for mediating the bottom-up and top-down processes. This not only represents the arrangement of fragmented regions for the bottom-up process, but also defines possible transformations of the object for the top-down process (Kimia, 2003).

One of many promising intermediate representations is *skeletonization*, where the skeleton of an object is extracted as a one-dimensional line representation, like the red line in Figure 9. The skeleton is formed by points that have at least two closest points on the object boundary (Cornea, Silver, & Min, 2007; Kimia, 2003). Green circles in Figure 9 illustrate that these points are centers of circles that are maximally inscribed within the boundary. Then, parts of the object are defined by skeleton branches, each of which is a line segment with no branch to multiple directions. Such parts can be consistently observed for different appearances of the object so they are useful for assembling fragmented regions in the bottom-up process. In addition, different configurations of parts represent various appearances of the object, and support the top-down process. Kimia (2003) presented the validity of skeletons from the psychophysical and neurophysiological perspectives. Also, researchers are exploring methods that recognize objects by appropriately matching parts of their skeletons (Bai & Latecki, 2008; Feinen, Yang, Tiebe, & Grzegorzek, 2014). Furthermore, these methods are being extended to realistic images with cluttered backgrounds, where an object is detected by applying contours of its parts to edges extracted for an image (Bai, Wang, Latecki, Liu, & Tu, 2009).

Traditional features are "hand-crafted" or "human-crafted" in the sense that their representations are specified in advance (Bengio, 2009). For instance, a SIFT feature is described as a 128-dimensional vector where each dimension represents the frequency of a certain edge orientation in a local region. However, such a hand-crafted feature is insufficient for representing diverse object appearances. This is because all of these appearances cannot be assumed in advance and cannot be appropriately represented by the feature. In the human brain, objects are recognized in a hierarchical fashion, where simple cells are gradually combined into more abstract, complex cells (Kruger et al., 2013). This hierarchical brain functionality is recently implemented as *deep learning* that constructs a feature hierarchy with higher-level features formed by the composition of lower-level features (Bengio, 2009; Bengio, Courville, & Vincent, 2013). Such a feature hierarchy is represented as a multilayer neural network. In every layer, each of the artificial neurons composes a more abstract feature based on outputs of neurons at the previous layer.

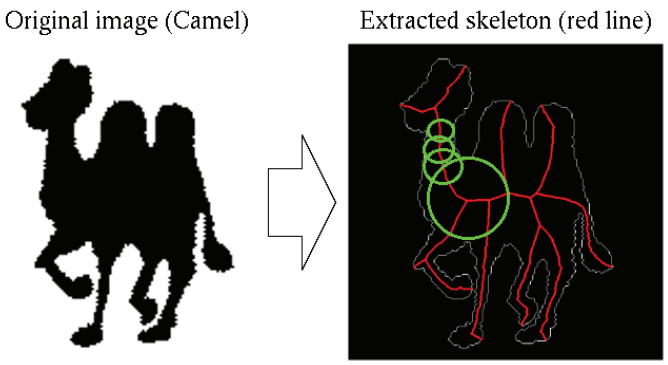

Figure 9. An illustration of skeletonization.

Figure 10 shows a conceptual comparison between a traditional machine learning approach using a hand-crafted feature and a deep learning approach. The former, in Figure 10 (a), uses a "shallow architecture" consisting of two layers, where the first layer transforms an example into a feature represented by a high-dimensional vector, and the second layer aggregates values of this feature into a detection result of a meaning. On the other hand, the deep learning in Figure 10 (b) first projects an example into the most primitive features at the bottom layer, and then these features are projected into more abstract ones at the second layer. This abstraction of features is iterated to obtain a detection result of a meaning. For examples, features at the bottom and second layers correspond to typical edges and their combinations, respectively. Moreover, features at an upper layer represent parts of a car, and the ones at the top layer indicate the whole car. Like this, the workflow from processing pixels to recognizing a meaning is unified into a deep architecture, which is extracted from large-scale data. One of the biggest advantages of this deep architecture is its discrimination power compared to the shallow one in the traditional machine learning approach. The latter requires $O(N)$ parameters to distinguish $O(N)$ examples, while the former can represent up to $O(2^N)$ examples using only $O(N)$ parameters (Bengio et al., 2013). Intuitively, a huge first layer is required for the traditional approach to discriminate diverse examples. In contrast, the discrimination power of the deep architecture is exponentially increased based on the combination of features at two consecutive layers.

For a long time, building such a deep architecture with a satisfying performance was difficult, but Hinton, Osindero, and Teh (2006) have developed an algorithm for reasonably solving this problem. The algorithm greedily builds one layer at a time so that outputs of the previous layer can be reconstructed with the minimal error rate. Using this as initialization, the deep architecture is finely tuned with training examples. In several worldwide competitions on image, video, and audio classification, the performance of deep learning

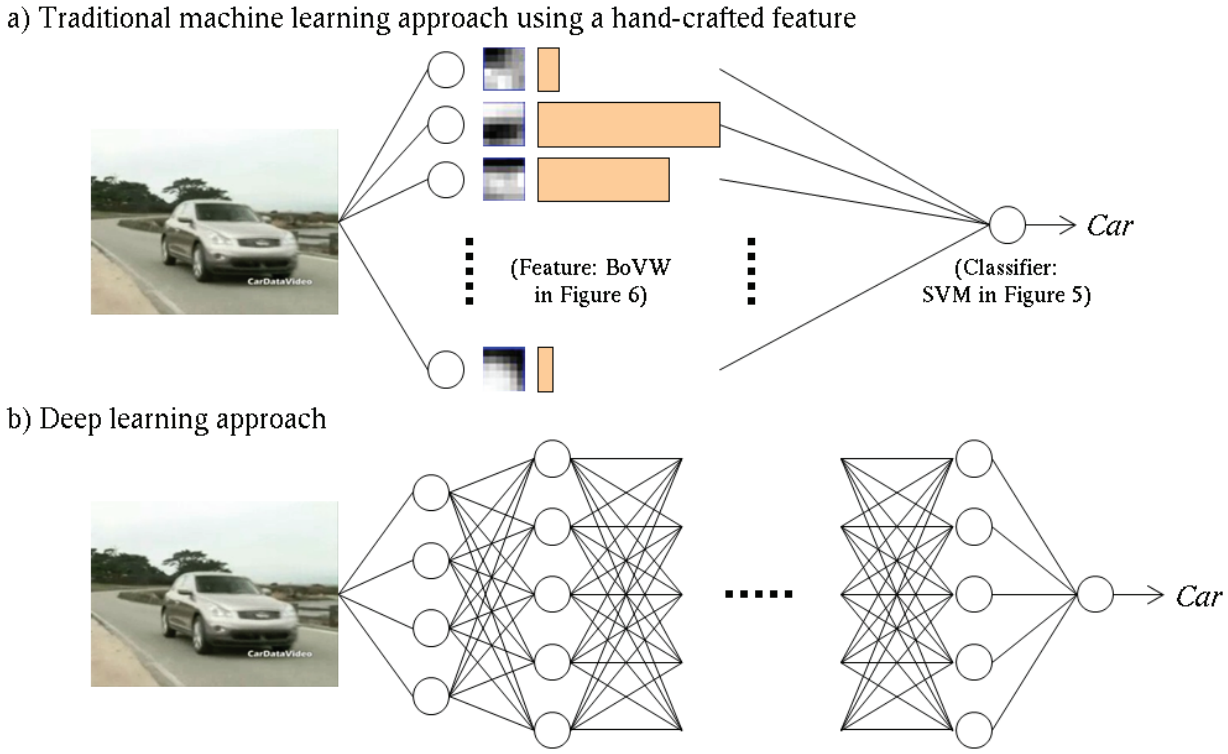

Figure 10. A conceptual comparison between traditional machine learning and deep learning approaches.

methods has been proven to be much higher than traditional machine learning methods (Krizhevsky, Sutskever, & Hinton, 2012; Lee, Pham, Largman, & Ng, 2009). Furthermore, in the field of neuroscience, it is well known that neurons encode sensory information using only 1–4% of active neurons (Bengio, 2009; Bengio et al., 2013). This idea is implemented as sparse coding and incorporated into deep learning by penalizing the output of each neuron.

Selective attention is the brain's mechanism that determines which part of sensory data is currently of most interest (Frintrop, Rome, & Christensen, 2010; Borji & Itti, 2013). This enables humans to conduct real-time decision-making by closely analyzing selected parts in a large amount of sensory data, such as sights and sounds captured by eyes and ears. *Visual attention* (also called focus of attention) implements selective attention on images and videos, that is, it detects salient regions that are likely to attract users (Frintrop et al., 2010; Borji & Itti, 2013). A detection result of such a salient region is usually represented as a *saliency map*, which represents the saliency of each pixel in an example. Figure 11 shows two examples of saliency maps where pixels associated with higher saliencies are depicted as brighter. It can be seen that the examples in Figure 11 (a) and (b) are appropriately associated with salient regions where a car and a person are shown, respectively. Since non-salient regions can be considered as irrelevant and redundant for interpreting semantic

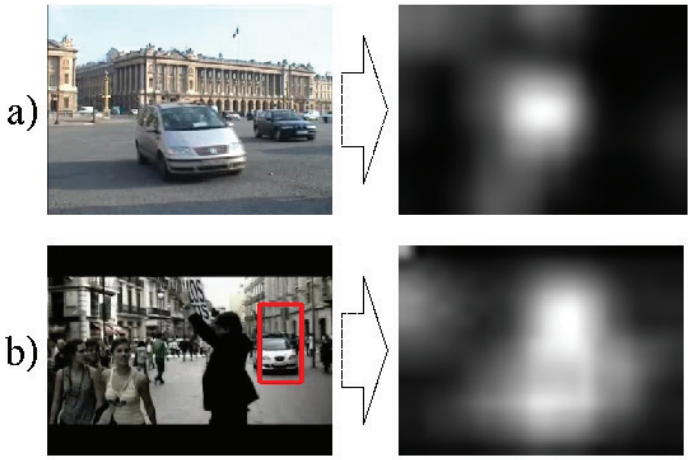

Figure 11. Examples of saliency maps.

meanings, visual attention is useful for not only improving the retrieval performance, but also reducing the computational cost.

In addition, visual attention facilitates analyzing the subjective property of examples. There is a big discrepancy between the goal of object recognition and that of retrieval. The former aims to recognize objects irrespective of various changing factors, such as directions, rotations, sizes, lighting conditions, and occlusion. However, this goal does not

fit user needs in retrieval. Let us consider a user retrieving examples where cars are shown. Clearly, he/she is not interested in an example where a car moves in a small background region (i.e., non-salient region), like the example in Figure 11 (b) in which the region of the car is marked by the red rectangle. Instead, an example where a car is shown in a salient region like the example in Figure 11 (a) should interest the user. Hence, visual attention is useful for evaluating the subjective property of each retrieved example and achieving meaningful retrieval for humans.

Most of the visual attention methods are based on the psychological theory called "feature integration theory," where different features (e.g., brightness, color contrast, and curvature) extracted for each pixel in an example are processed in parallel and fused into a saliency map (Borji & Itti, 2013; Frintrop et al., 2010). Typically, pixels which are irregular compared to surrounding ones are regarded as salient. However, this kind of bottom-up approaches relying only on features do not work well. Hence, researchers are exploring how to incorporate top-down approaches using prior knowledge into visual attention. For example, Li, Tian, Huang, and Gao (2010) proposed a method based on "contextual cueing," meaning that a human can easily find a target object when the visual context (i.e., spatial layout of objects) is similar to the past. This suggests that visual attention is guided by scenes that the human saw in the past. To implement contextual cueing, the method in Li et al. (2010) uses training examples where salient regions are labeled in advance. It detects salient regions in a test example by referring to the training example that has the most similar spatial layout.

Future Directions. Existing cognitive methods described above only utilize a small amount of knowledge ascertained in cognitive science. Below, we discuss further utilization of this knowledge in LSMR. One of the groundbreaking ideas that have emerged from the research on human categorization is the *prototype theory* (Lakoff, 1987; Mervis & Rosch, 1981; Rosch, 1975, 1978; Tversky, 1977). According to this theory, humans organize their concepts and categories into a radial structure centered around a prototype, with items closer to the prototype being deemed more central than those farther off. For instance, a pigeon is considered a more prototypical bird than a penguin. We can add to this the notion of "family resemblance" proposed by Wittgenstein (2009). The idea here is that members of a category have overlapping attributes, but there may be nothing that they all have in common.

Incorporating these features in an LSMR system requires that we are able to automatically cluster and label huge sets of images with large feature sets. This is an active area of research (Reed, Bifet, Holmes, Pfahringer, 2011; Spyromitros-Xioufis, Spiliopoulou, tsoumakas, & Vlahavas, 2011; Tsoumakas & Katakis, 2007), but we need to configure these techniques to produce structured clusters with different underlying similarity metrics, and design tools to explore and retrieve multimodal data from these clusters. (See, for instance, Koduri, Gali, and Indurkhya (2010) and Mala and Geetha (2009)).

Taking a different point of departure, Dastani and Indurkhya (1997) used Structural Information Theory and its notion of information load to introduce the measures of descriptor complexity and member complexity that drive categorization in opposite directions. They proposed a simple additive function to find an optimum balance between these two, and used it to model similarity and categorization. However, further research needs to be done to explore how these ideas scale up to huge databases.

As a pioneering work on automatic clustering/labeling of Internet scale data, Chen, Shrivastava, and Gupta (2013) developed NEIL (Never Ending Image Learner), which continuously explores those data to extract knowledge (positive images and relations for visual categories like objects, scenes, and attributes). First, for each category, seed images are collected through Google Image Search to build the initial classifier. Second, relations among categories are extracted by computing co-occurrences based on classifiers' outputs. Third, NEIL selects additional positive images, each of which has large outputs of both the classifier for a category and classifiers for its related categories. Then, NEIL updates classifiers with additional positive images and continuously repeats the second and third processes. As a result of running NEIL for 2.5 months, it could discover 400,000 positive examples and 1,700 relations for 2,237 categories. It seems possible to extend NEIL by adopting the prototype and structural information theories described above, so that more structured knowledge can be continuously extracted from Internet scale data.

Similarity, which is at the heart of LSMR, has been studied extensively from a cognitive science point of view (Goldstone & Son, 2005; Hahn, Chater, & Richardson, 2003; Rodriguez & Egenhofer, 2003; Schwering, 2008; Tversky, 1977). Thus, it would be useful to take advantage of some of these insights in designing LSMR systems. We cannot review here all the numerous cognitive studies on similarity, but we would like to make one comment on how they can help in LSMR. If we look at most of the existing formulations of similarity in LSMR systems, they are essentially feature based. In other words, certain features of the images are extracted, and then some similarity metric is applied on them. These features can be low-level perceptual features, or high-level semantic features. Needless to say, humans also use such features, but one distinguishing cognitive aspect of similarity is that it is highly dynamic and contextual. Moreover, depending on the context and the goal of the agent, new features can be created or discovered in an image that were not obvious or relevant before (Indurkhya, 1998; Indurkhya & Ojha, 2013). There has been some previous work in modeling these dynamic

processes (Hofstadter, 1995; O'Hara & Indurkhya, 1994), but we need to scale up these techniques, or come up with new techniques, so that they can be applied to huge databases.

In this regard, it would be useful to incorporate insights from the study of biological visual systems. Kruger et al. (2013), based on a thorough review of the existing literature on the primate visual system, have proposed three key mechanisms that need to be incorporated in computer vision systems:

1. Hierarchical processing: Features need to be grouped and organized in hierarchies. Moreover, these hierarchies need to be dynamic in the sense that they incorporate learning (with respect to both grouping and hierarchical structure), and are capable of evolving based on ongoing interactions with the environment.

2. Separate information channels depending on different needs for different behaviors or different requirements.

3. Feedback and feedforward: There should be both top-down and bottom-up mechanisms so that higher-level features can affect grouping of lower-level ones, and also lower-level features can evoke different higher-level ones.

Considering these key mechanisms, the current deep learning approach only implements the hierarchical processing of features. We expect that one important future direction for deep learning is to develop a mechanism that adapts (or projects) the feature hierarchy depending on images, so that high-level (semantic) features are consistently obtained in different situations (e.g., bright, dark, and foggy) where low-level perceptual features are dissimilar.

Also, the above kind of hierarchical architecture would be similar to the one proposed some years ago for modeling creativity in legal reasoning (Indurkhya, 1997). There is also more recent work on how perceptual and conceptual similarities interact together, and how perceptual similarities can give rise to new (hitherto unseen) conceptual similarities (Indurkhya & Ojha, 2013; Ojha & Indurkhya, 2009), which can be modeled with such an architecture.

ONTOLOGICAL APPROACHES

Existing Approaches. An *ontology* is a machine-readable representation of knowledge to explicitly specify concepts, properties of concepts, and relations among concepts in a given domain (Horridge, Knublauch, Rector, Stevens, & Roe, 2004; Staab & Struder, 2009). Concepts in multimedia data are defined as textual descriptions of semantic meanings that can be recognized by humans, such as objects like *Person* and *Car*, actions like *Walking* and *Airplane_Flying*, events like *Car_Crash* and *Explosion_Fire*, and scenes like *Beach* and *Desert*. Below, concept names are written in italics to distinguish them from the other terms. Ontological (also called concept-based) approaches have been developed where examples relevant to a query are retrieved based on detection results of concepts (Snoek & Worring, 2009).

Figure 12 illustrates an overview of an ontological approach based on the QBE framework in Figure 5. First of all, for each concept, a *detector* is built to detect its presence in an example. The detector associates the example with a detection score that represents a scoring value between 0 and 1. Large and small detection scores highlight the concept's presence and absence, respectively. For example, in Figure 12, the detector for *Person* provides the upper positive (user-provided) example with the score 0.9, meaning that a person probably appears in this example. On the other hand, the score 0.1, obtained by the detector for *Outdoor*, indicates that the upper positive example is unlikely to show an outdoor scene. By aggregating such detection scores for various concepts, an example is represented as a multidimensional vector and projected into the multidimensional space, as shown in the middle of Figure 12.

Given positive examples for a query (its text description can also be used as described in the classifier construction task below), a classifier is constructed in the multidimensional space of concept detection scores. Since the detector

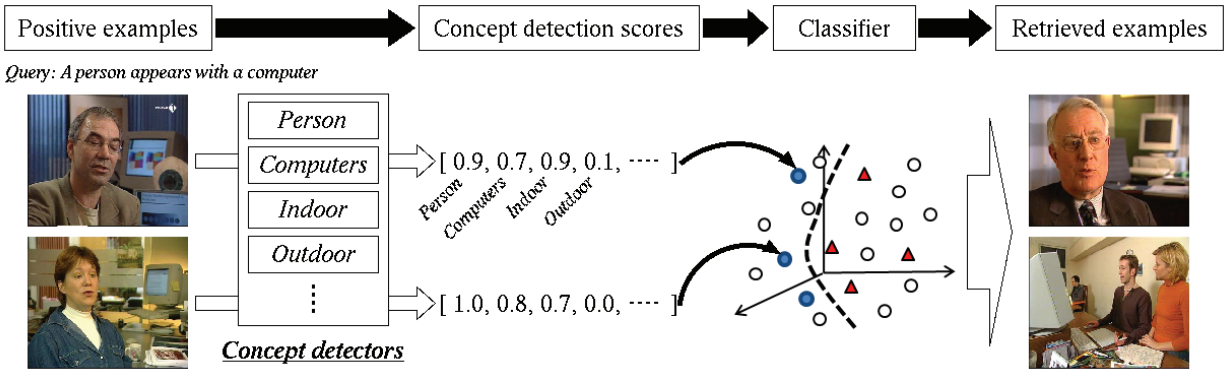

Figure 12. An overview of an ontological approach.

for each concept is built using a large amount of training examples, the concept can be robustly detected regardless of its sizes, positions, and directions on the screen. This enables us to collectively retrieve examples where concepts related to the query are present with diverse appearances. For example, positive examples in Figure 12 only show frontal views of *Computers*. But, as shown in the bottom right of Figure 12, the example showing the side view of *Computers* can be retrieved because the detector based on many training examples can assign high detection scores to examples showing different views of *Computers*. Please see Figure 5 where positive examples showing frontal views of *Computers* lead to only retrieve examples showing the same or very similar views. Like this, compared to the multidimensional space of features where each dimension just represents a physical characteristic of an example, ontological approaches take advantage of the space where each dimension represents the presence of a semantically meaningful concept. This facilitates retrieving examples that have dissimilar visual appearances, but show similar semantic meanings. This kind of ontological approaches achieve state-of-the-art retrieval performance (Li, Wang, Li, & Zhang, 2007; Natsev, Haubold, Těsić, Xie, & Yan, 2007; Ngo et al., 2009; Snoek et al., 2009; Wei, Jiang, & Ngo, 2011).

The following three tasks are crucial for ontological approaches. The first task is the question of how to define a vocabulary of concepts. Since a query is characterized by a set of concepts, a concept vocabulary should be sufficiently rich for covering various queries. One of the most popular ontologies is *Large-Scale Concept Ontology for Multimedia* (LSCOM), which defines a standardized set of 1,000 concepts in the broadcast news video domain (Naphade et al., 2006). These concepts are selected based on their "utility" for classifying content in videos, their "coverage" for responding to a variety of queries, their "feasibility" for automatic detection, and the "availability" (observability) of large-scale training data. It is estimated that if the number of concepts in LSCOM reaches an amount of 3,000, granting the quality of the new concepts remains similar to the existing ones, the retrieval performance approaches that of one of the best web search engines in text information retrieval (Hauptmann, Yan, Lin, Christel, & Wactlar, 2007). Apart from LSCOM, *ImageNet,* a large-scale ontology for images, is being developed (Deng et al., 2009). It is an extension to its predecessor, WordNet, which is a large lexical ontology where concepts (called synonym sets or synsets) are interlinked based on their meanings (Fellbaum, 1998). ImageNet aims to assign an average of 500 to 1,000 images to each WordNet concept. In Deng et al. (2009), 3.2 million images are associated with 5,247 concepts through Amazon's Mechanical Turk, where the assignment of images has been outsourced to web users (see the Manual Annotation Approaches section). The developers of ImageNet plan to assign 50 million images to 80,000 concepts in the near future.

The second task is figuring out how to accurately detect the presence of a concept in examples. It should be noted that concepts themselves are just linguistic terms. To utilize them in LSMR, we need to examine whether each concept is contained in the audiovisual form of an example. Hence, detectors serve as mediators between linguistic concepts and their audiovisual forms. As described in the Machine Learning Approaches section, much research effort has been made on developing accurate concept detectors (object recognizers) by mainly taking advantage of a large number of training examples and features exhaustively sampled in both the spatial and temporal dimensions. Concept detectors can be further improved by exploiting knowledge about the human visual system based on cognitive methods described in the previous section.

The last task concerns the utilization of detection scores to construct an accurate classifier for a query. This classifier fuses detection scores for multiple concepts into a single "relevance score," which indicates the relevance of an example to the query. Existing methods are roughly classified into four categories: *linear combination*, *discriminative*, *similarity-based*, or *probabilistic*. Linear combination computes the relevance score of an example by weighting detection scores for multiple concepts. One popular method is to use concept detection scores in positive examples. If the average detection score for a concept in positive examples is large, this concept is regarded as related to the query and associated with a large weight (Natsev et al., 2007; Wei et al., 2011). Another popular method is text-based weighting, where a concept is associated with a large weight if its name is lexically similar to a term in the text description of the query (Natsev et al., 2007; Wei et al., 2011). The lexical similarity between a concept name and a term can be measured using a lexical ontology like WordNet. Discriminative methods construct a classifier (typically, SVM) using positive examples (Natsev et al., 2007; Ngo et al., 2009) (see Figure 12). The relevance score of an example is obtained as the classifier's output. Similarity-based methods compute the relevance score of an example as the similarity between positive examples and the example in terms of concept detection scores. Li et al. (2007) use the cosine similarity and a modified entropy as similarity measures. Probabilistic methods estimate a probabilistic distribution of concepts using detection scores in positive examples, and use it to compute the relevance score of an example. Rasiwasia, Moreno, and Vasconcelos (2007) compute the relevance score of an example as the similarity between the multinomial distribution of concepts estimated from positive examples and the one estimated from the example.

Future Directions. Ontological approaches described above lack reasoning to precisely infer higher-level semantic meanings based on properties of concepts and their relations. Even though some works consider hierarchical relations

among concepts, they only use is-a (generalization/specialization) connections among concepts (Deng, Berg, & Fei-Fei, 2011; Zhu, Wei, & Ngo, 2013). Reasoning based on concept properties and relations is necessary because concept detection itself has the following two limitations. First, concepts are too general to identify examples that users want to retrieve. Secondly, most of the existing methods use concepts in isolation. For example, various semantic meanings are displayed in examples where the concepts *Person*, *Hand*, and *Ball* are present. In other words, examples that users really want cannot be identified by independently examining presences of *Person*, *Hand*, and *Ball*. Instead, if we consider that the *Hand* of a *Person* is moving and the *Ball* is separating from the *Person*, the higher-level meaning "throwing" can be derived.

Note that reasoning was explored in classical manual annotation approaches described in the Manual Annotation section. However, in LSMR, it has received little research attention due to the poor performance of concept detection in the past. Considering its recent improvement, we argue that reasoning should be addressed in LSMR. For this, Chen, Zhou, and Prasanna (2012) developed an interesting approach that optimally specializes detected concepts and their relations, so that they are the most probable and ontologically consistent. This approach, which formulates reasoning as an optimization problem based on constraints defined by the ontology, can be considered as a promising future direction of LSMR.

Reasoning requires overcoming the crucial problem of how to manage "uncertainties" in concept detection. Traditional ontology formalisms do not account for uncertainties, where an ontology itself is not uncertain. In other words, it is a presentation of prior knowledge that has been accepted to be true. Compared to this, even using the most effective detectors, it is still difficult to accurately detect various kinds of concepts. For example, our method, which performed the best at the concept detection competition in TRECVID 2012 (Shirahama & Uehara, 2012), can achieve high performances for concepts such as *Male_Person* and *Walking_Running* (with average precisions greater than 0.7). On the other hand, the detection of concepts like *Bicycling* and *Sitting_down* was difficult (with average precisions less than 0.1). In addition, real-world examples are "unconstrained" in the sense that they can be taken by arbitrary camera techniques and in arbitrary shooting environments (Jiang et al., 2013). Hence, even in the future, it cannot be expected to detect concepts with 100% accuracy. If one relies on uncertain concept detection results, detection errors for some concepts damage the whole reasoning process.

We have developed a method that can handle uncertainties based on *Dempster-Shafer Theory* (DST) (Shirahama, Kumabuchi, Grzegorzek, & Uehara, 2015). DST is a generalization of Bayesian theory where a probability is not assigned to a variable, but instead to a subset of variables (Denoeux, 2013). Given a set of concepts, C, and S, a subset of C, we define a "mass function" $m(S)$ over an example to indicate the probability that one concept in S is present in the example. For instance, $m(\{Person,Car\})$ represents the probability that either *Person* or *Car* could be present in an example. In the extreme case, $m(C)$ represents the probability that every concept could be present, that is, it is unknown which concept is present. Using such a mass function, DST can represent uncertainties in concept detection much more powerfully than Bayesian theory, because the latter can only represent uncertainties by assigning 0.5 to the probability of a concept's presence. However, the derivation of a mass function is quite intractable, because it is very subjective or impossible to prepare training examples by annotating them from the perspective that one of a set of concepts could be present. Thus, based on the set-theoretic operation in DST, we have proved that a probabilistic classifier using a mass function can be transformed into the one using "plausibilities." A plausibility is an upper bound probability that a concept could possibly be present in an example. By modeling these plausibilities based on the distribution of positive and negative examples for each concept, a classifier is constructed in the framework of maximum likelihood estimation. We have shown that this classifier yields about 19% performance improvement compared to a classifier that uses concept detection scores without considering uncertainties. One useful future direction might be to incorporate a reasoning mechanism into the above-mentioned classifier, where concept properties and relations are used as constraints in maximum likelihood estimation.

Furthermore, a large repository of concept properties and relations is required to reason various semantic meanings. In the text processing field, researchers are exploring *Information Extraction* (IE), which is the process of extracting relations between entities from natural language text (Alfonseca, Filippova, Delort, & Garrido, 2012). For example, the relation triples *Founding_location(University of Siegen, Germany)* and *Founding_year(University of Siegen, 1972)* are extracted from the sentence "University of Siegen in Germany was founded in 1972." By applying such an IE to multimedia data, we could create a large repository of concept properties and relations with or without a small amount of user intervention. We call this *Multimedia Information Extraction* (MIE) and consider it as a very important future direction. MIE can be considered as a "second generation" of video data mining described in the Heuristic Approaches section. Because of the poor performance of past concept detectors, video data mining could only analyze features (Shirahama et al., 2006). As a result, it failed to extract patterns characterizing high-level semantic meanings. MIE offers an opportunity to rethink video data mining by utilizing recent concept detectors that are much

more accurate than old ones. We have implemented a preliminary MIE system in which detection results for 351 concepts are probabilistically analyzed to extract higher-level meanings (Shirahama, Grzegorzek, & Uehara, 2015). We demonstrated that the high-level meaning *Birthday_Party* is appropriately characterized by concepts like *Moonlight*, *Nighttime*, *Entertainment*, *Singing*, and *Dancing*.

While our preliminary MIE system used concept detectors that merely identify the presence or absence of a concept, several detectors that can localize their regions are currently available (Felzenszwalb, Girshick, Mcallester, & Ramanan, 2010; Simonyan & Zisserman, 2014). Thus, we hope that MIE is further extended to consider spatio-temporal relations among concepts. For this, an example only displays the original 3D space, which is projected onto a 2D image plane. In other words, it does not hold the depth information in the original 3D space. For example, a 2D image or video frame may show that the regions of a *Person* and a *Table* are overlapping, even though the former stands in front of the latter. In addition, a *Ball* kicked hard and far by a *Football_Player* may still overlap with the player's 2D region. Compared to this, humans can easily interpret the depth information in 2D examples. This has inspired researchers to develop *depth estimation* methods, which estimate depths from 2D examples (Karsch, Liu, & Kang, 2012; Saxena, Chung, & Ng, 2008). Roughly speaking, some features are useful for predicting depths in an example: a grass field viewed at a short distance has fine textures, while such textures are blurred when it is viewed at a large distance. Furthermore, parallel lines have larger variations in edge orientations, as they are viewed from a more distant position. Based on such features, a classifier is built using training examples where the depth of each pixel is annotated (recorded) with a depth sensor like Microsoft Kinect. Intuitively, the classifier estimates depths in a test example by referring to those in visually similar training examples. We expect that depth estimation is necessary for MIE to analyze meaningful spatio-temporal relations among concepts.

ADAPTIVE APPROACHES

Existing Approaches. One way that a human gets to solve diverse problems is the repetition of the following process: Given a new problem, the human first monitors his/her performance, recognizes a deficiency, and uses knowledge that he/she already has to overcome the deficiency. By repeating this, the human can accumulate knowledge for solving diverse problems. In this context, *metacognition* is a discipline to explore the process of how a human addresses a problem (Anderson & Oates, 2007). Assuming a cognitive system that simulates a functionality of human mind, metacognition aims to monitor, model, and control the behavior of that system to effectively solve a problem. We define adaptive approaches as applications of metacognition to LSMR. The development of an LSMR system requires various decision-making capabilities, such as choosing a set of features, selecting a classifier, setting parameters, collecting training examples, selecting a performance evaluation measure, and so on. Adaptive approaches automate or optimize one or more decision-making tasks based on user feedback. This is an extension of Relevance Feedback (RF), described in the Interactive Approaches section.

The traditional RF relies on the very restrictive communication between a classifier and a user, where the user only informs the classifier whether an example is relevant to a certain semantic meaning or not. In the real world, a teacher makes much more complex communication with a learner. In particular, if the learner makes a mistake, the teacher tells him/her the reason for it. Based on this idea, Parkash and Parikh (2012) developed an *Attribute-based Feedback* (AF), which realizes the complex communication between a user and a classifier. Here, attributes are semantically meaningful descriptions, such as parts (e.g., "propeller"), shapes (e.g., "round"), textures (e.g., "stripe"), rough scene categories (e.g., "natural"), and nonverbal properties (e.g., "properties that dogs have but cats do not") (Farhadi, Endres, Hoiem, & Forsyth, 2009; Lampert, Nickisch, & Harmeling, 2009). Similar to concept detection, a detector for each attribute is built to identify its presence in an example. As a result, the example is represented as a vector, where each dimension represents the output of the detector for one attribute. For example, in Figure 13, the example (a) is associated with the large output value 0.6 for the attribute "natural" because trees and the grass are displayed in a large region. Note that the example representation based on attributes is similar to the one based on concepts (see Figure 12). But the attributes

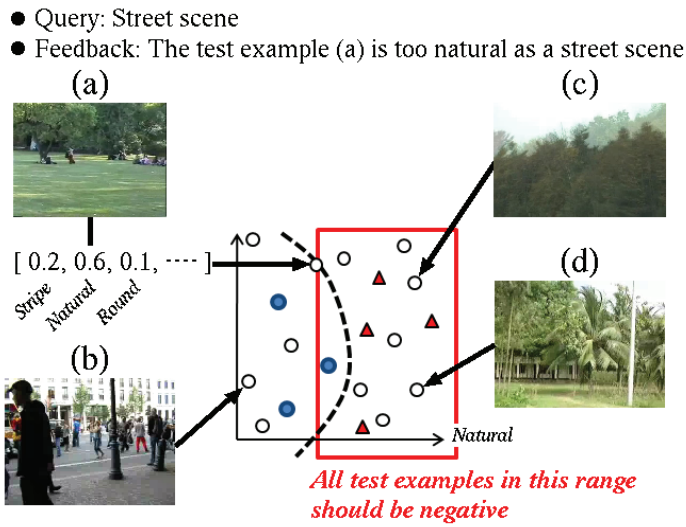

Figure 13. An overview of Attribute Feedback (AF).

represent lower-level semantic meanings, and are therefore relatively easier to detect automatically (Zhang et al., 2013).

AF uses attributes as a language between a classifier and a user to implement their complex communication (Parkash & Parikh, 2012). Specifically, if an example that the classifier regards as relevant to a meaning is judged to be irrelevant by a user, he/she can explain the reason for this misclassification. Let us consider Figure 13, where examples are represented as points in the multidimensional space defined by the detector outputs for different attributes. For simple visualization, only two dimensions are shown in Figure 13, where the horizontal dimension represents detector outputs for the attribute "natural." Assume that for the query "street scene," a classifier (SVM) with the boundary depicted by the dashed line is built using three positive and four negative examples, which are marked by blue circles and red triangles, respectively. Test examples are represented by white circles. Based on the criteria of RF, the user is asked to give feedback to the test example (a) because it is the closest to the classification boundary. Under this setting, the user can not only annotate the test example (a) as negative in terms of the query, but also explain "this example is too natural as a street scene." This implies that test examples that have higher detector outputs for the attribute "natural" than the test example (a) should be also negative. In Figure 13, these test examples like (c) and (d) are located in the red rectangle. Like this, based on the attribute explained in a reason, the annotation for one example can be propagated to other examples. That is, multiple examples are annotated through one feedback, so that the performance of a classifier can be effectively improved.

Furthermore, attributes, which are used as features of the classifier, can be refined based on user feedback (Biswas & Parikh, 2013). In Figure 13, the above exemplified explanation has another implication that the detector for the attribute "natural" should output lower values for positive examples than the one for the test example (a). Using this as a constraint, the detector is refined so that the positive example in the red rectangle is associated with a lower value than the one for the test example (a). This way, both the classifier and attributes (features) are refined by AF.

Future Directions. Adaptive approaches have plenty of room to explore. First, the current AF only targets the efficient refinement of classifiers for object-level meanings (concepts) based on attributes, but we expect that AF can be flexibly used for various levels of semantic meanings. Here, classifiers for a certain level of meanings are efficiently refined by regarding one lower level of meanings as attributes. In particular, AF seems to be useful for ontological approaches where concepts are considered as attributes, and accurate classifiers for high-level meanings can be built with reduced manual annotation effort. This is equivalent to effective knowledge extraction of MIE (see the Ontological Approaches section), because concept relations characterizing high-level meanings can be obtained by analyzing the built classifiers. Furthermore, by viewing these high-level meanings as attributes, AF may succeed in effortlessly extracting their causal relations, which were used in classical manual annotation approaches with huge manual labor.

Apart from AF, one important future direction for gaining the benefit from metacognition is to design *metalevel* features that are used to select an effective strategy for improving the retrieval performance. For example, Bensusan, Giraud-Carrier, and Kennedy (2000) suggested that the performance of a decision tree can be evaluated based on the number of nodes, depth, shape, and so on. Thus, using these as metalevel features, the decision tree that yields the best performance on given data can be estimated. In addition, Kumar, Packer, and Koller (2010) proposed "self-paced learning," which is inspired by the fact that children start with learning easier concepts, and then build up more complex ones. To implement this, the researchers developed a metalevel feature to assess the difficulty level of examples based on how easily their labels are predicted by the current classifier. From this, an accurate classifier can be constructed by gradually introducing training examples from easier to harder. We expect that various types of metalevel features are needed to characterize the usefulness of features, classifiers, and parameters in the LSMR processing pipeline.

Another major insight from metacognition is that humans conceptualize things in divergent ways. For example, while a frying pan is typically used for frying, it can also be used for hammering, fighting, or playing musical instruments. This kind of adaptive conceptualization in the human mind has been investigated as *gestalt projection* (Indurkhya, 2006; Koffka, 1935; Kubovy & Gepshtein, 2000). Gestalts are top-down structures that are used for modeling expectation-based approaches to how context affects the conceptualization of low-level sensory data. More specifically, we define gestalt projection as an extension of ontological approaches, and represent a gestalt as a structured set of concepts that are interrelated based on their postural, spatial, and temporal relations.

Let us consider that for the query "a person hammering," a user provides a positive example that shows "a person hammering a nail with a frying pan" (Guerin, Ferreira, & Indurkhya, 2014). In this case, the ontological approach in Figure 12 would retrieve examples having high detection scores for *Person* and *Frying_Pan*. However, this leads to retrieve many undesirable examples where a *Frying_Pan* is being used for cooking, is being washed, is being advertised, and so on. Thus, for accurate retrieval, we need to adaptively estimate that *Frying_Pan* in this positive example is being used for hammering. This "hammering" gestalt is evoked in the following way: For the positive example, the regions of *Person* and *Frying_Pan* are identified with the relational concept *Holding* (i.e., the former holds the latter). In addition, the

pose of *Person* suggests the action concept *Hitting* (which is not observed in most examples showing *Frying_Pan*). The above pattern of concepts and their interrelationships trigger the "hammering" gestalt. In this way, gestalt projection dynamically organizes the concepts detected in an example to yield the imaginative and playful conceptualization.

The following two tasks are the key to implementing the mechanism of gestalt projection. The first is to build a large-scale knowledge base about gestalts. This is exactly the task of Multimedia Information Extraction (MIE), discussed in the previous section. Furthermore, the computer vision community has started to develop methods that can identify group actions derived from the contextual relationships among multiple objects (Lan, Wang, Yang, Robinovitch, & Mori, 2012), expected social roles and actions of persons (Lan, Sigal, & Mori, 2012), and functionalities of objects (Zhu, Fathi, & Fei-Fei, 2014). These research efforts are beneficial to efficiently building a large-scale gestalt knowledge base.

The second task is to develop a method for applying an evoked gestalt to candidate examples. We feel that this does not require creating new technology, but rather to configure existing tools and mechanisms in new ways to bridge the semantic gap. One such platform might be the Blackboard System, which allows an interaction of bottom-up and top-down processes in a competition-cooperation paradigm to arrive at an interpretation of given perceptual data (Corkill, Lesser, & Hudlicka, 1982; Hayes-Roth, 1985). The blackboard architecture was originally proposed for speech understanding (Erman, Hayes-Roth, Lesser, & Reddy, 1980), but since then has been successfully applied in diverse domains (Corkill, 1991). This architecture may be visualized by the metaphor of a group of independent experts with diverse knowledge who are sharing a common workspace, namely the blackboard. They work on the solution together and each of them adds some contribution to the blackboard, whenever possible, until the problem is solved. The blackboard model provides an efficient platform for problems that require many diverse sources of knowledge. It allows a range of different experts represented as diverse computational agents and provides an integration framework for them. It enables an incremental progress toward a solution, and a flexible control for problem-solving. Integrating these two tasks in current LSMR technology would allow us to retrieve relevant examples to queries in an intuitive and humanlike way.

CONCLUSION

In this paper, we reviewed existing LSMR methods, including those that we have developed. By tracing the history of machine-based and human-based LSMR methods, we argued that due to prioritizing the generality of methods and the scalability for large-scale data, current methods lack knowledge about human interpretation, which was used in classical methods. We then discussed human-machine cooperation methods by classifying them into cognitive methods using knowledge about the human visual system, ontological methods using knowledge about human inference, and adaptive methods using knowledge about human learning. The future direction that we finally suggest is the development of a framework to unify cognitive, ontological, and adaptive methods into a single LSMR system by considering their relationships as shown in Figure 8. In this system, every process is based on knowledge about human interpretation of semantic meanings. We hope that this paper will be a trigger to disseminate the LSMR problem to other research fields and solve it in an interdisciplinary approach.

REFERENCES

Ahn, L. von, & Dabbish, L. (2004). Labeling images with a computer game. In E. Dykstra-Erickson & M. Tscheligi (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 319–326). New York, NY: ACM Press. http://dx.doi.org/10.1145/985692.985733

Ahn, L. von, & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51*(8), 58–67. http://dx.doi.org/10.1145/1378704.1378719

Ahn, L. von, Liu, R., & Blum, M. (2006). Peekaboom: A game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 55–64). New York, NY: ACM Press. http://dx.doi.org/10.1145/1124772.1124782

Alfonseca, E., Filippova, K., Delort, J.-Y., & Garrido, G. (2012). Pattern learning for relation extraction with a hierarchical topic model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papes—Volume 2* (54–59). Stroudsburg, PA: Association for Computational Linguistics.

Anderson, M. L., & Oates, T. (2007). A review of recent research in metareasoning and metalearning. *AI Magazine, 28*(1), 7–16. http://dx.doi.org/10.1609/aimag.v28i1.2025

Ando, R., Shinoda, K., Furui, S., & Mochizuki, T. (2006). Robust scene recognition using language models for scene contexts. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval* (pp. 99–106). New York, NY: ACM Press. http://dx.doi.org/10.1145/1178677.1178693

Ayache, S., & Quénot, G. (2008). Video corpus annotation using active learning. In *Proceedings of the 30th European Conference on IR Research* (pp. 187–198). Berlin, Germany: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-78646-7_19

Bai, X., & Latecki, L. J. (2008). Path similarity skeleton graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(7) 1282–1292. http://dx.doi.org/10.1109/TPAMI.2007.70769

Bai, X., Wang, X., Latecki, L. J., Liu, W., & Tu, Z. (2009). Active skeleton for non-rigid object detection. In *Proceedings of IEEE International Conference on Compuer Vision* (pp. 575–582).

Barrington, L., O'Malley, D., Turnbull, D., & Lanckriet, G. (2009). User-centered design of a social game to tag music. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 7–10). New York, NY: ACM Press. http://dx.doi.org/10.1145/1600150.1600152

Bell, M., Reeves, S., Brown, B., Sherwood, S., MacMillan, D., Ferguson, J., & Chalmers, M. (2009). EyeSpy: Supporting navigation through play. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 123–132). http://dx.doi.org/10.1145/1518701.1518723

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning, 2*(1), 1–127. http://dx.doi.org/10.1561/2200000006

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828. http://dx.doi.org/10.1109/TPAMI.2013.50

Bensusan, H., Giraud-Carrier, C. G., & Kennedy, C. J. (2000). A higher-order approach to meta-learning. In J. Cussens & A. Frisch (Eds.), *Proceedings of the Work-in-Progress Track at the 10th International Conference on Inductive Logic Programming* (pp. 43–59).

Bhatt, C., & Kankanhalli, M. (2011). Multimedia data mining: State of the art and challenges. *Multimedia Tools and Applications, 51*(1), 35–76. http://dx.doi.org/10.1007/s11042-010-0645-5

Biswas, A., & Parikh, D. (2013). Simultaneous active learning of classifiers & attributes via relative feedback. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 644–651). New York, NY: Institute of Electrical and Electronics Engineers. http://dx.doi.org/10.1109/CVPR.2013.89

Borji, A., and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207. http://dx.doi.org/10.1109/TPAMI.2012.89

Chang, P., Han, M., & Gong, Y. (2002). Extract highlights from baseball game video with hidden Markov models. In *Proceedings of the IEEE International Conference on Image Processing* (pp. 609–612). New York, NY: Institute of Electrical and Electronics Engineers.

Chen, N., Zhou, Q.-Y., & Prasanna, V. (2012). Understanding web images by object relation network. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 291–300). New York, NY: ACM Press.

Chen, X., Shrivastava, A., & Gupta, A. (2013). NEIL: Extracting visual knowledge from web data. In *Proceedings of the 2013 IEEE Internationl Conference on Computer Vision (ICCV)* (pp. 1409–1416). New York, NY: Institute of Electrical and Electronics Engineers.

Corkill, D. D. (1991). Blackboard systems. *AI Expert, 6*(9), 40–47.

Corkill, D. D., Lesser, V. R., & Hudlicka, E. (1982). Unifying data-directed and goal-directed control. In *Proceedings of the Second National Conference on Artificial Intelligence*, 143–147. Pao Alto, CA: Association for the Advancement of Artificial Intelligence.

Cornea, N. D., Silver, D., & Min, P. (2007). Curve-skeleton properties, applications, and algorithms. *IEEE Transactions on Visualization and Computer Graphics, 13*(3), 530–548. http://dx.doi.org/10.1109/TVCG.2007.1002

Csurka, G., Bray, C., Dance, C., Fan, L., & Williamowski, J. (2004). Visual categorization with bags of keypoints. In *Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision* (pp. 1–22).

Dastani, M., & Indurkhya, B. (1997). An algebraic approach to similarity and categorization. In *Proceedings of SIMCAT 1997: An Interdisciplinary Workshop on Similarity and Categorisation* (pp. 51–57).

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Survey, 40*(2), 5:1–5:60. http://dx.doi.org/10.1145/1348246.1348248

Deng, J., Berg, A. C., & Fei-Fei, L. (2011). Hierarchical semantic indexing for large scale image retrieval. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 785–792). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/CVPR.2011.5995516

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/CVPR.2009.5206848

Denoeux, T. (2013). Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering, 25*(1), 119–130. http://dx.doi.org/10.1109/TKDE.2011.201

Djordjevic, D., Izquierdo, E., & Grzegorzek, M. (2007). User driven systems to bridge the semantic gap. In *Proceedings of the 15th European Signal Processin Conference* (pp. 718–722).

Erman, L. D., Hayes-Roth, F., Lesser, V. R., & Reddy, D. R. (1980). The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys, 12*(2), 213–253. http://dx.doi.org/10.1145/356810.356816

Fan, R.-E., Chen, P.-H., & Lin, C.-J. (2005). Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research, 6,* 1889–1918.

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1778–1785). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/CVPR.2009.5206772

Feinen, C., Yang, C., Tiebe, O., & Grzegorzek, M. (2014). Shape matching using point context and contour segments. In D. Cremers, I. Reid, H. Saito, & M.-H. Yang (Eds.), *the 12th Asian Conference on Computer Vision* (pp. 95–110). Switzerland: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-16817-3_7

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1627–1645. http://dx.doi.org/10.1109/TPAMI.2009.167

François, A. R. J., Nevatia, R., Hobbs, J., Bolles, R. C., & Smith, J. R. (2005). VERL: An ontology framework for representing and annotating video events. *IEEE MultiMedia, 12*(4), 76–86. http://dx.doi.org/10.1109/MMUL.2005.87

Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception, 7*(1), 6:1–6:39. http://dx.doi.org/10.1145/1658349.1658355

Gemmell, D. J., Vin, H. M., Kandlur, D. D., Rangan, P. V., & Rowe, L. A. (1995). Multimedia storage servers: A tutorial. *IEEE Computer, 28*(5), 40–49. http://dx.doi.org/10.1109/2.384117

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. Holyoak & R. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 13–36). Cambridge, MA: MIT Press.

Grauman, K., & Liebe, B. (2011). *Visual object recognition (synthesis lectures on artificial intelligence and machine learning)*. San Rafael, CA: Morgan & Claypool Publishers.

Guerin, F., Ferreira, P. A., & Indurkhya, B. (2014). Using analogy to transfer manipulation skills. In *Proceedings of the Twenty-Eighth Conference on Artificial Intelligence* (pp. 14–19). Pao Alto, CA: Association for the Advancement of Artificial Intelligence.

Gupta, M., Li, R., Yin, Z., & Han, J. (2010). Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter, 12*(1), 58–72. http://dx.doi.org/10.1145/1882471.1882480

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition, 87*(1), 1–32. http://dx.doi.org/10.1016/S0010-0277(02)00184-1

Hauptmann, A., Yan, R., Lin, W.-H., Christel, M., & Wactlar, H. (2007). Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE Transactions on Multimedia, 9*(5), 958–966. http://dx.doi.org/10.1109/TMM.2007.900150

Hayes-Roth, B. (1985). A blackboard architecture for control. *Artificial Intelligence, 26*(3), 251–321. http://dx.doi.org/10.1016/0004-3702(85)90063-3

Hinton, G. E., Osindero, S., & The, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*(7), 1527–1554. http://dx.doi.org/10.1162/neco.2006.18.7.1527

Hofstadter, D. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York, NY: Basic Books.

Horridge, M., Knublauch, H., Rector, A., Stevens, R., & Wroe, C. (2004). *A practical guide to building OWL ontologies with the Protege-OWL plugin edition 1.0*. Manchester, UK: University of Manchester. Retrieved from http://www.researchgate.net/publication/230585369_A_Practical_Guide_To_Building_OWL_Ontologies_Using_The_Protg-OWL_Plugin_and_CO-ODE_Tools

Indurkhya, B. (1997). On modeling creativity in legal reasoning. In *Proceedings of the 6th International Conference on Artificial Intelligence and Law* (pp. 180–189). New York, NY: ACM Press. http://dx.doi.org/10.1145/261618.261651

Indurkhya, B. (1998). On creation of features and change of representation. *Journal of Japanese Cognitive Science Society, 5*(2), 43–56.

Indurkhya, B. (2006). Emergent representations, interaction theory, and the cognitive force of metaphor. *New Ideas in Psychology, 24*(2), 133–162. http://dx.doi.org/10.1016/j.newideapsych.2006.07.004

Indurkhya, B., & Ojha, A. (2013). An experimental study on the role of perceptual similarity in visual metaphor. *Metaphor and Symbol, 28*(4), 233–253.

Inoue, N., & Shinoda, K. (2012). A fast and accurate video semantic-indexing system using fast MAP adaptation and GMM supervectors. *IEEE Transactions on Multimedia, 14*(4), 1196–1205. http://dx.doi.org/10.1109/TMM.2012.2191395

Izquierdo, E., Chandramouli, K., Grzegorzek, M., & Piatrik, T. (2007). K-Space content management and retrieval system. In *Proceedings of the 14th International Conference on Image Analysis and Processing Workshops* (pp. 131–136). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/ICIAPW.2007.32

Jain, A.K., Vailaya, A., & Wei, X. (1999). Query by video clip. *Multimedia Systems*, 7(5), 369–384. http://dx.doi.org/10.1007/s005300050139

Jiang, Y.-G., Yang, J., Ngo, C.-W., & Hauptmann, A. G. (2010). Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions*

on Multimedia, 12(1), 42–53. http://dx.doi.org/10.1109/TMM.2009.2036235

Jiang, Y.-G., Bhattacharya, S., Chang, S.-F., & Shah, M. (2013). High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval, 2*(2), 73–101. http://dx.doi.org/10.1007/s13735-012-0024-2

Juszczak, P., & Duin, R. P. W. (2003). Selective sampling methods in one-class classification problems. In *Proceedings of the 2003 International Conference on Artificial Neural Networks and Neural Information Processing* (pp. 140–148). Berlin: Springer-Verlag.

Karsch, K., Liu, C., & Kang, S. B. (2012). Depth extraction from video using non-parametric sampling. In *Proceedings of the 12th European Conference on Computer Vision—Volume Part V* (pp. 775–788). Berlin: Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-33715-4_56

Kashino, K., Kurozumi, T., & Murase, H. (2003). A quick search method for audio and video signals based on histogram pruning. *IEEE Transactions on Multimedia, 5*(3), 348–357. http://dx.doi.org/10.1109/TMM.2003.813281

Kim, Y.-T., & Chua, T.-S. (2005). Retrieval of news video using video sequence matching. In *Proceedings of the 11th International Multimedia Modeling Conference* (pp. 68–75). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/MMMC.2005.63

Kimia, B. B. (2003). On the role of medial geometry in human vision. *Journal of Physiology-Paris, 97*(2–3), 155–190.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 453–456). New York, NY: ACM Press. http://dx.doi.org/10.1145/1357054.1357127

Koduri, G. K., Gali, A., & Indurkhya, B. (2010). REM: A ray exploration model that caters to the search needs of multi-attribute data. In *Proceedings of the 2010 ACM Workshop on Social, Adaptive, and Personalized Multimedia Interaction and Access Pages* (pp. 49–54). New York, NY: ACM Press. http://dx.doi.org/10.1145/1878061.1878078

Koffka, K. (1935). *Principles of Gestalt psychology*. London, UK: Routledge.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems* (pp. 1106–1114).

Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., . . . Wiskott, L. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1847–1871. http://dx.doi.org/10.1109/TPAMI.2012.272

Kubovy, M., & Gepshtein, S. (2000). Gestalt: From phenomena to laws. In K. Boyer & S. Sarkar (Eds.), *Perceptual organisation for artificial vision systemsi* (pp. 42–72). Dordrecht, The Netherlands: Kluwer Academic.

Kumar, M. P., Packer, B., & Koller, D. (2010). Self-paced learning for latent variable models. In *Proceedings of Advances in Neural Information Processing Systems* (pp. 1189–1197).

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. Chicago, IL: Univeristy of Chicago Press.

Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 951–958). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/CVPR.2009.5206594

Lan, T., Sigal, L., & Mori, G. (2012). Social roles in hierarchical models for human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Rrecognition* (pp. 1354–1361). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/CVPR.2012.6247821

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., & Mori, G. (2012). Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(8), 1549–1562. http://dx.doi.org/10.1109/TPAMI.2011.228

Lee, H., Pham, P., Largman, Y., & Ng, A. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Proceedings of Advances in Neural Information Processing Systems* (pp. 1096–1104).

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications, 2*(1), 1–19. http://dx.doi.org/10.1145/1126004.1126005

Li, J., Tian, Y., Huang, T., & Gao, W. (2010). Probabilistic multi-task learning for visual saliency estimation in video. *International Journal of Computer Vision, 90*(2), 150–165. http://dx.doi.org/10.1007/s11263-010-0354-6

Li, X., & Snoek, C. G. M. (2009). Visual categorization with negative examples for free. In *Proceedings of the 17th ACM International Conference on Multimedia* (pp. 661–664). New York, NY: ACM Press. http://dx.doi.org/10.1145/1631272.1631382

Li, X., Wang, D., Li, J., & Zhang, B. (2007). Video search in concept subspace: A text-like paradigm. In *Proceedings of the 6th International Conference on Image and Video Retrieval* (pp. 603–610). New York, NY: ACM Press. http://dx.doi.org/10.1145/1282280.1282366

Liu, X., Zhuang, Y., & Pan, Y. (1999). A new approach to retrieve video by example video clip. In *Proceedings of*

the 7th International Conference on Multimedia (Part 2) (pp. 41–44). New York, NY: ACM Press. http://dx.doi.org/10.1145/319878.319889

Liu, Y., Zhang, D., Lu, G., & Ma, W. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition, 40*(1), 262–282. http://dx.doi.org/10.1016/j.patcog.2006.04.045

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the 7th International Conference on Computer Vision* (pp. 1150–1157). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/ICCV.1999.790410

Mala, T., & Geetha, T. V. (2009). Multi-level categorization visualization based on gestalt perception model. *Gestalt Theory, 31*(1), 43–54.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89–113. http://dx.doi.org/10.1146/annurev.ps.32.020181.000513

Monaco, J. (1981). *How to read a film*. New York, NY: Oxford University Press.

Naphade, M., Smith, J. R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., . . . Curtis, J. (2006). Large-scale concept ontology for multimedia. *IEEE MultiMedia, 13*(3), 86–91. http://dx.doi.org/10.1109/MMUL.2006.63

Naphade, M. R., & Smith, J. R. (2004). On the detection of semantic concepts at TRECVID. In *Proceedings of the 12th International Conference on Multimedia* (pp. 660–667). New York, NY: ACM Press. http://dx.doi.org/10.1145/1027527.1027680

Natsev, A., Naphade, M. R., & Tešić, J. (2005). Learning the semantics of multimedia queries and concepts from a small number of examples. In *Proceedings of the 13th International Conference on Multimedia* (pp. 598–607). New York, NY: ACM Press. http://dx.doi.org/10.1145/1101149.1101288

Natsev, A., Haubold, A., Těsić, J., Xie, L., & Yan, R. (2007). Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th International Conference on Multimedia* (pp. 991–1000). New York, NY: ACM Press. http://dx.doi.org/10.1145/1291233.1291448

Ngo, C., Jiang, Y.-G., Wei, X.-Y., Zhao, W., Liu, Y., Wang, J., . . . Chang, S.-F. (2009). VIREO/DVM at TRECVID 2009: High-level feature extraction, automatic video search and content-based copy detection. In *Proceedings of TRECVID 2009* (pp. 415–432).

Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *Proceedings of the 9th European Conference on Computer Vision* (pp. 490–503). Berlin: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/11744085_38

Ogiela, M. R., & Tadeusiewicz, R. (2010). Towards new classes of cognitive vision systems. In *Proceedings of the IEEE International Conference on Complex, Intelligent, Software Intensive Systems* (pp. 851–855). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/CISIS.2010.49

O'Hara, S., & Indurkhya, B. (1994). Incorporating (re)-interpretation in case-based reasoning. In S. Wess, K.-D. Althoff, & M. M. Richter (Eds.), *Topics in case-based reasoning* (pp. 246–260). Berlin: Springer Berlin Heidelberg.

Ojha, A., & Indurkhya, B. (2009). Perceptual vs. conceptual similarities and creation of new features in visual metaphor. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *New frontiers in analogy research* (pp. 348–357). Sofia, Bulgaria: New Bulgarian University Press.

Oomoto, E., & Tanaka, K. (1993). OVID: Design and implementation of a video-object database system. *IEEE Transactions on Knowledge and Data Engineering, 5*(4), 629–643. http://dx.doi.org/10.1109/69.234775

Parkash, A., & Parikh, D. (2012). Attributes for classifier feedback. In *Proceedings of the 12th European Conference on Computer Vision* (pp. 354–368). Berlin: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-33712-3_26

Pattanasri, N., Chatvichienchai, S., & Tanaka, K. (2005). Towards a unified framework for context-preserving video retrieval and summarization. In *Proceedings of the 8th International Conference on Asian Digital Libraries* (pp. 119–128). Berlin: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/11599517_14

Peng, Y., & Ngo, C.-W. (2005). EMD-based video clip retrieval by many-to-many matching. In *Proceedings of the 4th Conference on Image and Video Retrieval* (pp. 71–81). http://dx.doi.org/10.1007/11526346_11

Petkovic, M., & Jonker, W. (2002). *Content-based video retrieval: a database perspective*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pizlo, Z. (2010). *3D Shape: Its unique place in visual perception*. Cambridge, MA: The MIT Press.

Pizlo, Z. (2014). *Making a machine that sees like us*. New York, NY: Oxford University Press.

Quinn, A. J., & Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1403–1412). New York, NY: ACM Press. http://dx.doi.org/10.1145/1978942.1979148

Rasiwasia, N., Moreno, P. J., & Vasconcelos, N. (2007). Bridging the gap: query by semantic example. *IEEE Transactions on Multimedia, 9*(5), 923–938. http://dx.doi.org/10.1109/TMM.2007.900138

Reed, J., Bifet, A., Holmes, G., & Pfahringer, B. (2011). Streaming multi-label classification. In *Proceedings of the Second Workshop on Applications of Pattern Analysis, Volume 17* (pp. 19–25).

Rodriguez, A., & Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different

ontologies. *IEEE Transactions on Knowledge and Data Engineering, 15*(2), 442–456. http://dx.doi.org/10.1109/TKDE.2003.1185844

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General, 104*(3), 192–233. http://dx.doi.org/doi/10.1037/0096-3445.104.3.192

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Mahwah, NJ: Lawrence Erlbaum Associates.

Russakovsky, O., Ma, S., Krause, J., Deng, J., Berg, A., & Li, F.-F. (2014). ImageNet large scale visual recognition challenge. Retreieved from http://www.image-net.org/challenges/LSVRC/2014/

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1–3), 157–173. http://dx.doi.org/10.1007/s11263-007-0090-8

Sande, K. E. A. van de, Gevers, T., & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1582–1596. http://dx.doi.org/10.1109/TPAMI.2009.154

Sande, K. E. A. van de, Gevers, T., & Snoek, C. G. M. (2011). Empowering visual categorization with the GPU. *IEEE Transactions on Multimedia, 13*(1), 60–70. http://dx.doi.org/10.1109/TMM.2010.2091400

Saxena, A., Chung, S. H., & Ng, A. Y. (2008). 3-D depth reconstruction from a single still image. *International Journal of Computer Vision, 76*(1), 53–69. http://dx.doi.org/10.1007/s11263-007-0071-y

Schmid, C., & Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(5), 530–535. http://dx.doi.org/10.1109/34.589215

Schwering, A. (2008). Approaches to semantic similarity measurement for geo-spatial data: A survey. *Transactions in GIS, 12*(1), 5–29. http://dx.doi.org/10.1111/j.1467-9671.2008.01084.x

Shirahama, K., & Grzegorzek, M. (2014). Towards large-scale multimedia retrieval enriched by knowledge about human interpretation. *Multimedia Tools and Applications*. http://dx.doi.org/10.1007/s11042-014-2292-8

Shirahama, K., Grzegorzek, M., & Uehara, K. (2015). Weakly supervised detection of video events using hidden conditional random fields. *International Journal of Multimedia Information Retrieval, 4*(1), 17–32. http://dx.doi.org/10.1007/s13735-014-0068-6

Shirahama, K., Ideno, K., & Uehara, K. (2006). A time-constrained sequential pattern mining for extracting semantic events in videos. In V.A. Petrushin & L. Khan (Eds.), *Multimedia data mining and knowledge discovery* (pp. 404–426). London, UK: Springer. http://dx.doi.org/10.1007/978-1-84628-799-2_20

Shirahama, K., Matsuoka, Y., & Uehara, K. (2012). Event retrieval in video archives using rough set theory and partially supervised learning. *Multimedia Tools and Applications, 57*(1), 145–173. http://dx.doi.org/10.1007/s11042-011-0727-z

Shirahama, K., Otaka, K., & Uehara, K. (2007). Content-based video retrieval using video ontology. In *Proceedings of the 3rd International Workshop on Multimedia Information Processing and Retrieval* (pp. 283–289).

Shirahama, K., & Uehara, K. (2012). Kobe University and Muroran Institute of Technology at TRECVID 2012 semantic indexing task. In *Proceedings of TRECVID 2012* (pp. 239–247).

Shirahama, K., Kumabuchi, K., Grzegorzek, M., & Uehara, K. (2015). Video retrieval based on uncertain concept detection using Dempster-Shafer theory. In A. K. Baughman, J. Gao, J.-Y. Pan, & V. Petrushin (Eds.), *Multimedia data mining and analytics* (pp. 269–294). Switzerland: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-14998-1_12

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349–1380. http://dx.doi.org/10.1109/34.895972

Snoek, C. G. M., van de Sande, K. E. A., de Rooij, O., Huurnink, B., Uijlings, J. R. R., van Liempt, M., . . . Smeulders, A. W. M. (2009). The MediaMill TRECVID 2009 semantic video search engine. In *Proceedings of TRECVID 2009* (pp. 226–238).

Snoek, C. G. M., & Smeulder, A. W. M. (2012). Internet video search. Tutorial on ECCV 2012. Retrieved from http://tutorials.ceessnoek.info/

Snoek, C. G. M., & Worring, M. (2009). Concept-based video retrieval. *Foundations and Trends in Information Retrieval, 2*(4), 215–322. http://dx.doi.org/10.1561/1500000014

Snoek, C. G. M., Worring, M., Geusebroek, J.-M., Koelma, D., & Seinstra, F. J. (2005). On the surplus value of semantic video analysis beyond the key frame. In *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 386–389). New York, NY: Institute of Electrical Engineers. http://dx.doi.org/10.1109/ICME.2005.1521441

Spyromitros-Xioufis, E., Spiliopoulou, M., Tsoumakas, G., & Vlahavas, I. (2011). Dealing with concept drift and class imbalance in multi-label stream classification. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (pp. 1583–1588). Palo Alto, CA: Association

for the Advancement of Artificial Intelligence. http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-266

Staab, S., & Studer, R. (Eds.). (2009). *Handbook on ontologies* (2nd ed.). Springer-Verlag Berlin Heidelberg.

Staab, S., Scherp, A., Arndt, R., Troncy, R., Grzegorzek, M., Saathoff, C., . . . Hardman, L. (2008). Semantic Multimedia. In C. Baroglio, P. A. Bonatti, J. Maluszyński, M. Marchiori, A. Polleres, & S. Schaffert (Eds.), *Reasoning web* (pp. 125–170). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-85658-0_4

Steggink, J., & Snoek, C. G. M. (2011). Adding semantics to image-region annotations with the Name-It-Game. *Multimedia Systems, 17*(5), 367–378. http://dx.doi.org/10.1007/s00530-010-0220-y

Syrett, M. (2009, March 24). Number of tags per video on YouTube [Blog post]. *Independent Films by the Numbers*. Retrieved from http://www.lathrios.com/blog/?p=70

Tanaka, K., Ariki, Y., & Uehara, K. (1999). Organization and retrieval of video data. *IEICE Transactions on Information and Systems, 82*(1), 34–44.

Tong, S., & Chang, E. (2001). Support vector machine active learning for image retrieval. In *Proceedings of the 9th International Conference on Multimedia* (pp. 107–118). New York, NY: ACM Press. http://dx.doi.org/10.1145/500141.500159

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining, 3*(3), 1–13. http://dx.doi.org/10.4018/jdwm.2007070101

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352. http://dx.doi.org/doi/10.1037/0033-295X.84.4.327

Uehara, K., Oe, M., & Maehara, K. (1996). Knowledge representation, concept acquisition and retrieval of video data. In *Proceedings of the International Symposium on Cooperative Database Systems for Advanced Applications* (pp. 527–534).

Vapnik, V. (1998). *Statistical learning theory*. New York, NY: Wiley.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 511–518). New York, NY: Institute for Electrical Engineers. http://dx.doi.org/10.1109/CVPR.2001.990517

Volkmer, T., Smith, J. R., & Natsev, A. (2005). A web-based system for collaborative annotation of large image and video collections: An evaluation and user study. In *Proceedings of the 13th International Conference on Multimedia* (pp. 892–901). http://dx.doi.org/10.1145/1101149.1101341

Wang, M., & Hua, X.-S. (2011). Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology, 2*(2), 10:1–10:21. http://dx.doi.org/10.1145/1899412.1899414

Wei, X.-Y., Jiang, Y.-G., & Ngo, C.-W. (2011). Concept-driven multi-modality fusion for video search. *IEEE Transactions on Circuits and Systems for Video Technology, 21*(1), 62–73. http://dx.doi.org/10.1109/TCSVT.2011.2105597

Wittgenstein, L. (2009). *Philosophical investigations*. New York, NY: Blackwell Publishing.

Yan, R., & Hauptmann, A. G. (2007). A review of text and image retrieval approaches for broadcast news video. *Information Retrieval, 10*(4–5), 445–484. http://dx.doi.org/10.1007/s10791-007-9031-y

Yan, R., Fleury, M.-O., Merler, M., Natsev, A., & Smith, J. R. (2009). Large-scale multimedia semantic concept modeling using robust subspace bagging and MapReduce. In *Proceedings of the 1st ACM Workshop on Large-Scale Multimedia Retrieval and Mining* (pp. 35–42). New York, NY: http://dx.doi.org/10.1145/1631058.1631067

Yoshitaka, A., Ishii, T., Hirakawa, M., & Ichikawa, T. (1997). Content-based retrieval of video data by the grammar of film. In *Proceedings of the IEEE Symposium on Visual Languages* (pp. 310–317). New York, NY: Institute for Electrical Engineers. http://dx.doi.org/10.1109/VL.1997.626599

YouTube. (n.d.). Statistics. *YouTube*. Retrieved from http://www.youtube.com/yt/press/statistics.html

Yuan, J., Tian, Q., & Ranganath, S. (2004). Fast and robust search method for short video clips from large video collection. In *Proceedings of the 17th International Conference on Pattern Recognition (Volume 3)* (pp. 866–869). New York, NY: Institute for Electrical Engineers. http://dx.doi.org/10.1109/ICPR.2004.1334665

Zha, Z.-J., Yang, L., Mei, T., Wang, M., Wang, Z., Chua, T.-S., & Hua, X.-S. (2010). Visual query suggestion: Towards capturing user intent in internet image search. *ACM Transactions on Multimedia Computing,Communications, and Applications, 6*(3), 13:1–13:19. http://dx.doi.org/10.1145/1823746.1823747

Zhai, Y., Rasheed, Z., & Shah, M. (2004). A framework for semantic classification of scenes using finite state machines. In *Proceedings of the 3rd International Conference on Image and Video Retrieval* (pp. 279–288). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-27814-6_35

Zhang, H., Zha, Z.-J., Yang, Y., Yan, S., Gao, Y, & Chua, T.-S. (2013). Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st International Conference on Multimedia* (pp. 33–42). New York, NY: ACM Press. http://dx.doi.org/10.1145/2502081.2502093

Zhang, J., Marszalek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision, 73*(2), 213–238. http://dx.doi.org/10.1007/s11263-006-9794-4

Zhou, X. S., & Huang, T. S. (2003). Relevance feedback in image retrieval: A comprehensive review. *Multimedia*

Systems, 8(6), 536–544. http://dx.doi.org/10.1007/s00530-002-0070-3

Zhu, S., Wei, X.-Y., & Ngo, C.-W. (2013). Error recovered hierarchical classification. In *Proceedings of the 21st International Conference on Multimedia* (pp. 697–700). New York, NY: ACM Press. http://dx.doi.org/10.1145/2502081.2502182

Zhu, Y., Fathi, A., & Fei-Fei, L. (2014). Reasoning about object affordances in a knowledge base representation. *Proceedings of the 13th European Conference on Computer Vision* (408–424). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-10605-2_27

Zwol, R. von, Garcia, L., Ramirez, G., Sigurbjornsson, B., & Labad, M. (2008). Video tag game. In *Proceedings of the 17th International World Wide Web Conference*. New York, NY: ACM Press.